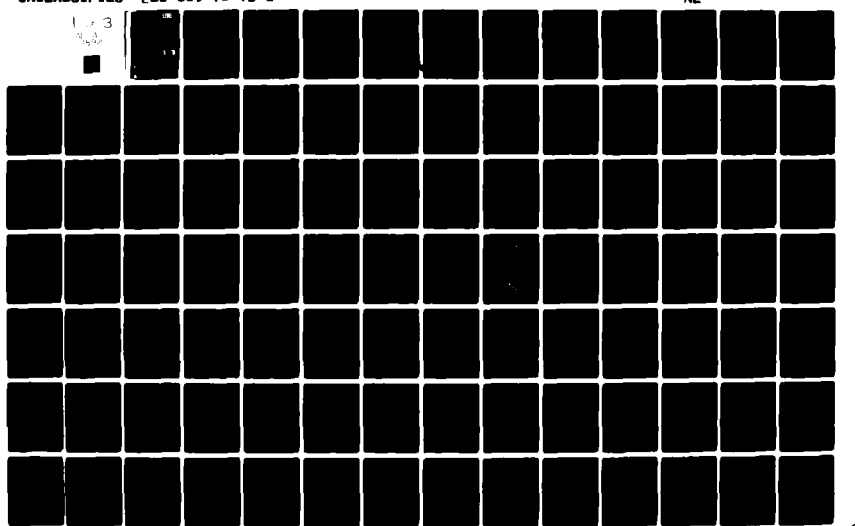
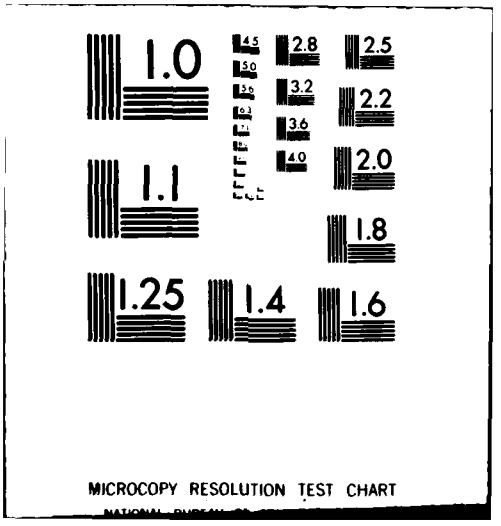
AD-A089 210 GEORGIA INST OF TECH ATLANTA SCHOOL OF ELECTRICAL EN--ETC F/G 17/2
AN ANALYSIS OF OBJECTIVE MEASURES FOR USER ACCEPTANCE OF VOICE --ETC(U)
SEP 79 T P BARNWELL, W D VOIERS DCA100-78-C-0003
UNCLASSIFIED E21-659-78-TB-1 NL




1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
6.3
7.1
8.0
9.0
MICROCOPY RESOLUTION TEST CHART

DCA 100-78-C-0003

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER (14) E21-659-78-TB-1	2. GOVT ACCESSION NO. AD-A089 320	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) (6) An Analysis of Objective Measures for User Acceptance of Voice Communication Systems,		5. TYPE OF REPORT & PERIOD COVERED (9) Final Report,
		6. PERFORMING ORG. REPORT NUMBER N/A
7. AUTHOR(s) (10) Thomas P. Barnwell, III; William D. Voiers		8. CONTRACT OR GRANT NUMBER(s) (15) DCA100-78-C-0003
9. PERFORMING ORGANIZATION NAME AND ADDRESS Georgia Institute of Technology School of Electrical Engineering Atlanta, Georgia 30332 (12) 222		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS (11) Sep 79
11. CONTROLLING OFFICE NAME AND ADDRESS Defense Communications Engineering Center 1860 Wiehle Ave, (R540) Reston, VA 22090		12. REPORT DATE September 1979
		13. NUMBER OF PAGES 230
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) 408632		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE N/A

16. DISTRIBUTION STATEMENT (of this Report)

Unlimited, Open Publication

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

SAME

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Speech, Speech Quality, Quality, Objective Testing, Subjective Testing, Voice Communications, Speech Communications, Coding, Speech Coding

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The report presents the results of a large study of the statistical correlation between a data base of subjective speech quality measures and a data base of objective speech quality measures. Both data bases were derived from approximately eighteen hours of coded and distorted speech. The subjective test used was the Diagnostic Acceptability Measure (DAM) test ~~developed at the Dynastat Corporation~~. The objective measures included spectral distance measures, frequency variant spectral distance measures, signal-to-noise measurements,

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

408632

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

area ratio distance measures, log area ratio distance measures, PARCOR distance measures, log PARCOR distance measures, feedback coefficient distance measures, log feedback coefficient distance measures, residual energy ratio distance measures, and composite measures. The analysis procedures included linear regression analysis, multiple linear regression analysis, and nonlinear regression analysis. In all, approximately 1,500 variations of these objective measures were studied.

The figure-of-merit used for measuring the performance of an objective measure was the estimated correlation coefficient between the objective measure and the subjective data base. Parametrically different distance measures were compared using nonparametric pairwise rank statistics.

The results of this study give quantitative predictions of the performance of many objective speech quality measures for predicting subjective user acceptance. Further, this study forms a basis for choosing among parametrically different forms of the same objective distance measure.

AN ANALYSIS OF OBJECTIVE MEASURES FOR USER ACCEPTANCE OF VOICE COMMUNICATIONS SYSTEMS

by

Thomas P. Barnwell III

School of Electrical Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332

and

William D. Voiers

Dynastat, Inc.
2704 Rio Grande
Austin, Texas 78705

FINAL REPORT

DCA 100-78-C-0003

Prepared For

Defense Communications Agency
Defense Communications Engineering Center
1860 Wiehle Avenue
Reston, Virginia 22090

September 1979

TABLE OF CONTENTS

		Page
1	INTRODUCTION	1
	1.1 Task History	1
	1.2 Technical Background	1
	1.3 An Approach to Designing and Testing Objective Quality Measures	5
	1.4 Principal Goals and Procedures	9
	1.5 Summary of Major Results	12
	1.6 Discussion	14
	References	17
2	SUBJECTIVE CRITERIA OF SPEECH ACCEPTABILITY	18
	2.1 Background	18
	2.2 Design of the Diagnostic Acceptability Measure (DAM)	18
	2.3 Materials and Procedures	26
	2.3.1 Speech Materials	26
	2.3.2 Evaluation Procedures	26
	2.3.3 Listener Selection and Calibration	27
	2.3.4 Analysis of DAM Data	28
	References	32
3	OBJECTIVE MEASURES	3
	3.1 Introduction	
	3.2 Basic Concepts and Notations	
	3.3 The Simple Measures	
	3.3.1 The Spectral Distance Measures	
	3.3.1.1 The LPC Parametric Analysis Techn	
	3.3.1.2 The Computation of Objective Measu	
	3.3.2 Parametric Distance Measures	
	3.3.3 Simple Noise Measurements	
	3.4 Frequency Variant Objective Measures	
	3.4.1 Banded Spectral Distance Measures	
	3.4.2 Banded Noise Measures	
	3.5 The Composite Measures	
	References	
4	THE DISTORTED DATA BASE	
	4.1 The Coding Distortions	
	4.1.1 Simple Waveform Coders	
	4.1.1.1 CVSD	
	4.1.1.2 ADM	
	4.1.1.3 APCM	
	4.1.1.4 ADPCM	
	4.1.2 The LPC Vocoder	
	4.1.3 The Adaptive Predictive Coder (APC)	77
	4.1.4 The Voice Excited Vocoder (VEV)	80
	4.1.5 Adaptive Transform Coding (ATC)	84

Table of Contents (Continued)

		Page
4.2	The Controlled Distortions	86
4.2.1	Simple Controlled Distortions	86
4.2.1.1	Additive Noise	86
4.2.1.2	Filtering Distortions	88
4.2.1.3	Interruptions	88
4.2.1.4	Clipping	88
4.2.1.5	Center Clipping	92
4.2.1.6	Quantization Distortion	94
4.2.1.7	Echo Distortion	94
4.2.2	Frequency Variant Controlled Distortions	94
4.2.2.1	Additive Colored Noise	96
4.2.2.2	The Pole Distortion	96
4.2.2.3	The Banded Frequency Distortion	100
References		104
5	EFFECTS OF SELECTED FORMS OF DEGRADATION ON SPEECH ACCEPTABILITY AND ITS PERCEPTUAL CORRELATES	105
5.A	Methods and Materials	106
5.A.A	Listening Crews	106
5.A.B	Speakers	106
5.B	Experimental Results	107
5.1	Degradation by Coding	107
5.1.1	Simple Waveform Coders	108
5.1.1.1	Effects of Continuously Variable-Slope Delta Modulation (CVSD) on DAM Scores	108
5.1.1.2	Effects of Adaptive Delta Modulation (ADM) on DAM Scores	110
5.1.1.3	Effects of Adaptive Pulse Code Modulation APCM) on Dam Scores	110
5.1.1.4	Effects of Adaptive Differential Pulse Code Modulation (ADPCM) on DAM Scores	110
5.1.2	The Effects of Linear Predictive Coding on DAM Scores	114
5.1.3	The Effects of Adaptive Predictive Coding on DAM Scores	114
5.1.4	Effects of Voice Excited Vocoding (VEV) on DAM Scores	115
5.2	Controlled Degradation	115
5.2.1	Simple Forms of Controlled Degradation	115
5.2.1.1	Effects of Additive Broad-Band Noise on DAM Scores	120
5.2.1.2	Effects of Frequency Distortion on DAM Scores	122
5.2.1.2-1	Effects of bandpass filtering on DAM scores	123
5.2.1.2-2	Effects of low-pass filtering on DAM scores	123
5.2.1.2-3	Effects of high-pass filtering on DAM scores	126

Table of Contents (Continued)

		Page
5.2.1.3	Effects of Periodic Interruption on DAM Scores	126
5.2.1.4	Effects of Peak Clipping on DAM Scores	128
5.2.1.5	Effects of Center Clipping	132
5.2.1.6	Effects of Signal Quantization on DAM Scores	134
5.2.1.7	Effects of Echo on DAM Scores	134
5.2.2	Effects of Frequency-Variant Controlled Distortion on DAM Scores	136
5.2.2.1	Effects of Additive Colored Noise on DAM Scores	136
5.2.2.2	Effects of Pole Distortion on DAM Scores	138
5.2.2.2-1	Effects of pole frequency distortion	138
5.2.2.2-2	Effects of radial pole distortion	142
5.2.2.3	Banded Frequency Distortion	145
References		148
6	THE EXPERIMENTAL RESULTS	149
6.1	Introduction	149
6.2	Analysis Procedures	149
6.2.1	The Estimation Procedures	150
6.2.2	The Distorted Data Sets	153
6.2.3	The Subjective Data Sets	155
6.2.4	Non-parametric Rank Statistics	155
6.3	The Spectral Distance Measure Results	158
6.3.1	The Best Spectral Distance Measures	160
6.3.2	The Effect of Energy Weighting	162
6.3.3	The Effects of Spectral Weighting	162
6.3.4	The Effects of L_p Averaging	165
6.3.5	The Effect of the Pointwise Nonlinearity	168
6.3.6	The Effects of Other Subjective Measures	168
6.3.7	The Effects of Different Distorted Data Bases	171
6.3.8	The Effects of Nonlinear Regression Analysis	171
6.4	Simple Noise Measures	174
6.5	The Parametric Distance Measures	176
6.5.1	The Best Parametric Distance Measures	176
6.5.2	The Log Area Ratio Measure	178
6.5.3	The Energy Ratio Distance Measure	189
6.6	Frequency Variant Measures	189
6.6.1	The Frequency Variant Spectral Distance Measures	189
6.6.2	Frequency Variant Noise Measurements	199
6.7	The Composite Distance Measures	204
6.7.1	The Composite Measure Used to Measure Mutual Information	205
6.7.2	Composite Measures for Maximum Correlation	208
References		210

LIST OF FIGURES

		Page
1.2-1	System for Computing Objective Quality Measures	3
1.3-1	Block Diagram for System for Comparing the Effectiveness of Objective Quality Measures	7
2.2-1	DAM Rating Forms	22
3.3.1-1	Comparison of Fourier and LPC Spectra for a Vowel . . .	37
3.3.2-1	Computation of the Residual Energy Distance Measure . .	48
3.3.3-1	System for Computing Short Time SNR	50
3.4.2-1	Computation of Components of Short Time Banded Signal-Noise Measurements for the n^{th} frame and B channels. The filters, F1-FB, are non-overlapping band-pass filters.	58
4.1.1-1	General System for Describing Waveform Coders	67
4.1.2-1	Linear Productive Coder (LPC) Simulated to Form the LPC Coding Distortion	75
4.1.3-1	The Adaptive Predictive Coding System Used As Part of the Coding Distortion Study	79
4.1.4-1	Voiced Excited Vocoder	82
4.1.5-1	Adaptive Transform Coder Used for the Distorted Data Base	85
4.2.2.1-1	System for Creating the Frequency Variant Additive Noise Distortion	97
4.2.2.2-1	System for Producing the Frequency Variant Pole Distortions	99
4.2.2.3-1	System for Implementing the Banded Frequency Distortion	102
5.1.1.1	Effects of Continuously-variable Slope Delta Modulation on DAM Scores for Male and Female Speakers .	109
5.1.1.2	Effects of Adaptive Delta Modulation on DAM Scores for Male and Female Speakers	111
5.1.1.3	Effects of Adaptive Differential Pulse Code Modulation DAM Scores for Male and Female Speakers . .	112
5.1.1.4	Effects of Adaptive Differential Pulse Code Modulation on DAM Scores for Male and Female Speakers .	113
5.1.2	Effects of Linear Predictive Coding on DAM Scores for Male and Female Speakers	116
5.1.3	Effects of Adaptive Predictive Coding on DAM Scores for Male and Female Speakers	117
5.1.4.1	Effects of Voice-excited Vocoding (7 level Quantization) on DAM Scores for Male and Female Speakers	118
5.1.4.2	Effects of Voice-excited Vocoding (13 level Quantization) on DAM Scores for Male and Female Speakers	119
5.2.1.1	Effects of Broad-band Guassian Noise on DAM Scores for Male and Female Speakers	121

List of Figures (Continued)

		Page
5.2.1.2-1	Effects of Band-pass Filtering on DAM Scores for Male and Female Speakers	124
5.2.1.2-2	Effects of Low-pass Filtering on DAM Scores for Male and Female Speakers	125
5.2.1.2-3	Effects of High-pass Filtering on DAM Scores for Male and Female Speakers	127
5.2.1.3-1	Effects of Rapid Periodic Interruption on DAM Scores for Male and Female Speakers	129
5.2.1.3-2	Effects of Slower Periodic Interruption on DAM Scores for Male and Female Speakers	130
5.2.1.4	Effects of Peak-Clipping on DAM Scores for Male and Female Speakers	131
5.2.1.5	Effects of Center-clipping on DAM Scores for Male and Female Speakers	133
5.2.1.6	Effects of Quantization on DAM Scores for Male and Female Speakers	135
5.2.2.1M	Effects of Narrow-band noise on DAM Scores for Male Speakers	137
5.2.2.1F	Effects of Narrow-band Noise on DAM Scores for A Female Speaker	139
5.2.2.2-1M	Effects of Pole-Frequency Distortion on DAM Scores for Male Speakers	140
5.2.2.2-1F	Effects of Pole-frequency Distortion on DAM Scores for Female Speaker	141
5.2.2.2-2M	Effects of Radial Pole Distortion on DAM Scores for Male Speakers	143
5.2.2.2-2F	Effects of Radial Pole Distortion on DAM Scores for Female Speakers	144
5.2.2.3M	Effects of Banded Frequency Distortion on DAM Scores for Male Speakers	146
5.2.2-3F	Effects of Banded Frequency Distortion on DAM Scores for a Female Speaker	147

LIST OF TABLES

		Page
1.4-1	Summary of the Objective Quality Measures Studied . . .	11
2.2-1	Structure of the DAM	24
4-1	Total Set of Distortions in the Distorted Data Base	64
4-2	Contents of the Individual DAM Runs	65
4.1.1.1-1	Parameters for CVSD	69
4.1.1.2-1	Parameters for Adaptive Delta Modulator (ADM)	71
4.1.1.3-1	Parameters for Adaptive Pulse Code Modulation (APCM)	73
4.1.1.4-1	Parameters for Adaptive Differential Pulse Code Modulation (ADPCM)	74
4.1.2-1	Parameters for the LPC Vocoder	78
4.1.3-1	Parameters for the Adaptive Predictive Coder (APC)	81
4.1.4-1	Parameters for the Voice Excited Vocoder (VEV)	83
4.1.5-1	Parameters for the Adaptive Transform Coder	87
4.2.1.1-1	The Additive Noise Distortion	89
4.2.1.2-1	Filter Characteristics for Recursive Filters Used for Filter Distortion	90
4.2.1.3-1	"Keep" and "Drop" Constants for Intercept Distortion	91
4.2.1.4-1	Clipping Constants for Clipping Distortion	93
4.2.1.5-1	Center Clipping Constant for Center Clipping Distortion	93
4.2.1.6-1	Quantization Distortion Parameters	95
4.2.1.7-1	Echo Constant for the Echo Distortion	95
4.2.2.1-1	Colored Noise Distortions	98
4.2.2.2-1	Pole Distortion Control Parameters	101
4.2.2.3-1	Control Parameters for Banded Noise Distortion	103
6.2.2-1	Subclasses of Distortions Used as Part of this Research	154
6.2.4-1	Example Layout for the Results of a Four Parameter Paired Ranking Test	157
6.3-1	Summary of the 192 Spectral Distance Measures Studied	159
6.3.1-1	Best Five Spectral Distance Measures for CA, TSQ, and TBQ Across ALL and WBD	161
6.3.2	Rank Test Results for Energy Weighting	163
6.3.3-1	Rank Test Results for Spectral Weighting by $V(m,p,d,\phi)$ for Spectral Distance Measures	164
6.3.4-1(a)	Rank Test Results for L_p Norm for Spectral Distance Measures	166
6.3.4-1(b)	Rank Test Results for L_p Norm for Spectral Distance Measures	167
6.3.5-1	Pairwise Rank Test for δ on the $\|\|^{\delta}$ Nonlinearity Plus the Log Nonlinearity	169

List of Tables (Continued)

		Page
6.3.6-1	Maximum Correlation over all Spectral Distance Measures for Different Subjective Measures	170
6.3.7-1	Maximum Correlation Values for Spectral Distance Measures for CA, TSQ, and TBQ over the Different Subsets of the Distorted Data Base	172
6.3.8-1	The Effects of Non-Linear Regression Analysis on Spectral Distance Measures. Only maximum results are shown.	173
6.4-1	Results for SNR and Short Time SNR for CA Across WFC and ND	175
6.5-1	Summary of Parameters for Parametric Distance Measures	177
6.5.1-1	Best Six Results for Linear Feedback Parametric Distance Measure	179
6.5.1-2	Best Six Results for Log PARCOR Parameter Distance Measure	180
6.5.1-3	Best Six Results for Log Feedback Coefficient Parametric Distance Measure	181
6.5.1-4	Best Six Results for Linear Area Ratio Parametric Distance Measure	182
6.5.1-5	Six Best Results for the Linear PARCOR Parametric Distance Measure	183
6.5.1-6	Best Six Results for Log Area Ratio Parametric Distance	184
6.5.1-7	Best Six Results for the Energy Ratio Parametric Distance Measure	185
6.5.2-1	Total Results for Log Area Ratio Parametric Measure for CA, TSQ, and TBQ for ALL and WBD	186
6.5.2-2	The Maximum Values for CA for the Log Area Ratio Measure Across Different Distortion Subsets	187
6.5.2-3	The Effects of Higher Order Regression Analysis on the Log Area Ratio Distance Measure	188
6.5.3-1	Maximum Results from the Energy Ratio Distance Measure	190
6.5.3-2	The Maximum Value of CA for the Energy Ratio Measure Across Different Distortion Subset	191
6.5.3-3	The Effects of Higher Order Regression Analysis on the Energy Ratio Distance Measure	192
6.6-1	Frequency Bands Used for the Frequency Variant Objective Measures	193
6.6.1-1	Summary of 96 Frequency Variant Spectral Distance Measures Tested	195
6.6.1-2	Best Five Systems for Each Category for Log Frequency Variant Spectral Distance Measures	196
6.6.1-3	Best Five Systems for Each Category for Linear Frequency Variant Spectral Distance Measures	197
6.6.1-4	Sample of Results for Frequency Variant Spectral Distance Measures Used for Predicting Parametric Subjective Results	198

List of Tables (Continued)

		Page
6.6.2-1	Summary of 49 Short Time Banded Signal-to-Noise Ratio Measure	200
6.6.2-2	Best Five Results for Banded Short Time SNR Measure Across WFC	201
6.6.2-3	Results of the Pairwise Ranking Test for the Energy Weighting Parameter, α, for the Short Time Signal-to-Noise Ratio	202
6.6.2-4	Results of the Pairwise Ranking Test for the Power Parameter α for the Banded Short Time Signal-to-Noise Ratio	203
6.7.1-1	Results of the Composite Distance Measure Tests to Measure Mutual Information Among Different Distance Measures	206
6.7.2-1	The Best Composite Measures Discovered During this Study	209

CHAPTER 1

INTRODUCTION

1.1 Task History

The research effort reported here was performed jointly by the School of Electrical Engineering of the Georgia Institute of Technology and the Dynastat Corporation for the Defense Communications Agency. In this effort, the Georgia Institute of Technology was the prime contractor and the Dynastat Corporation was the subcontractor. The monitoring officer at the Defense Communications Engineering Center was originally Dr. William Bellfield. The monitoring officer was later changed to be Mr. James Vest.

This task, the investigation of the correlation between objective and subjective measures for speech quality, followed previous work by both Georgia Tech [1.1] and the Dynastat Corp. [1.2] [1.3] in related areas. The portion of this research performed at Georgia Tech involved the production of distorted and coded speech, the measurement of objective quality measures, and the correlation of the objective measures with the subjective measures. The portion of the work performed at the Dynastat Corp. included subjective quality testing and the associated analysis.

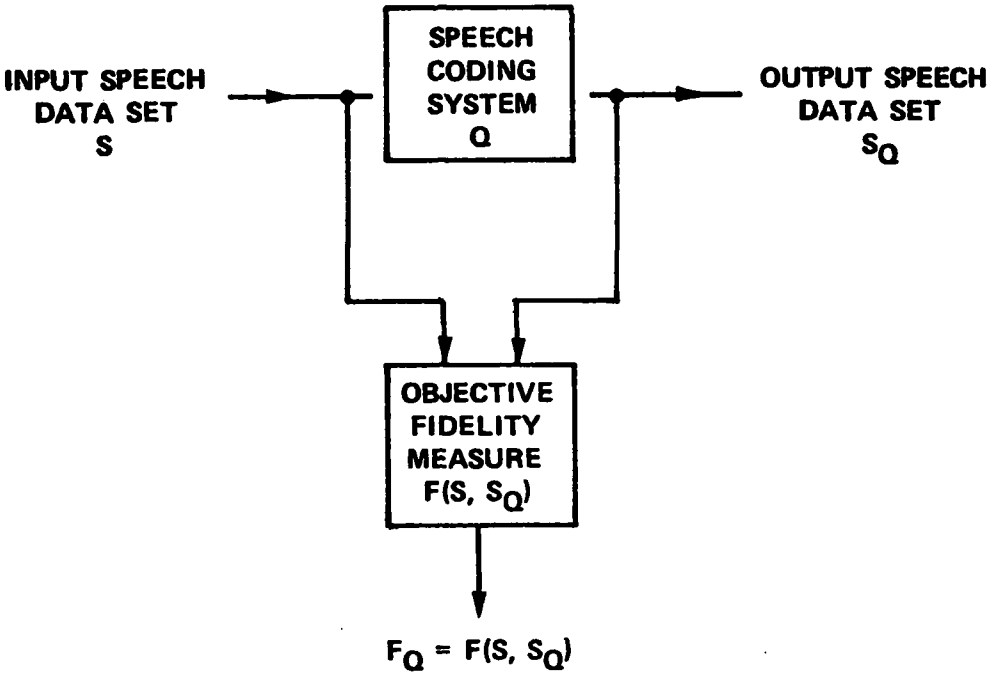
1.2 Technical Background

Since it has been clear for some years that some form of end-to-end speech digitization would be initiated by the Defense Communications Systems, a number of speech digitization systems have been developed at various laboratories around the country. The job of selecting from these candidate systems the features to be included in the final system requires

that extensive evaluation and testing be performed. Likewise, when a "final" system is fielded, periodic and initial field testing of all links will be a significant requirement. This effort deals with a set of techniques which can be used for more effective and efficient operational speech quality testing. In general, these "objective fidelity measures" are computed from an "input" or "unprocessed" speech data set, S, and an "output" or "distorted" speech data set, S_Q, as shown in Figure 1.2-1. The output speech data set results when the input speech data set is passed through the speech communication system under test. Objective measures may be very simple, such as the traditional signal-to-noise ratio, or they may be very complex. A complex measure might use such diverse measures as a spectral distance or other parameteric distances between the input and output speech data sets; semantic, syntactic, or phonemic information extracted from the input speech data set; or the characteristics or the talker's vocal tract or glottis. If an objective fidelity measure conforms to the triangular inequality and the other conditions shown in Figure 1.2-1, then it is a metric. Although metrics have many features which are desirable in a fidelity measure, an objective measure need not be metric to be of interest.

If an objective fidelity measure existed which was both highly correlated with the results of human preference tests and which was also compactly computable, then its utility would be undeniable. Clearly, it could be used instead of subjective quality measures for testing and optimizing speech coding systems. Such tests could be expected to be less expensive to administer, to give more consistent results, and, in general, not to be subject to the human failings of administrator or subject. Such an objective measure would also be very useful in the design of speech

OBJECTIVE FIDELITY MEASURES



CONDITIONS FOR A MEASURE TO BE A METRIC

1. $F(S, S_Q) = F(S_Q, S)$
2. $F(S, S_Q) = 0$ if $S = S_Q$
   $F(S, S_Q) \geq 0$ if $S \neq S_Q$
3. $F(S, S_Q) \leq F(S, S_Y) + F(S_Y, S_Q)$

Figure 1.2-1. System for Computing Objective Quality Measures.

coding systems, either by iterative optimization of the parameters of the coding system by repeatedly applying the quality measure--a process which is extremely expensive using subjective tests--or, if the procedure were analytically tractable, by designing the speech coding system to explicitly maximize the quality of the system as defined by the objective quality measure. Finally, note that the results of the objective measure applied at different times and at different locations could be compared directly. This is clearly not generally the case for the results of subjective quality tests.

The problem is that an objective fidelity measure which is both highly correlated with subjective measures over all possible distortions, and which is compactly computable, does not exist. Although at this time the speech perception process is not well understood, it is well enough understood to state that the human speech perceiver is an active perceiver, responding to semantic, syntactic, and talker related information as well as phonemic content, and that he uses his vast knowledge of the language interactively in the speech perception process. The acoustic correlates of the various hierarchically structured elements of the language in the speech signal are simultaneously overlapping and redundant. This means that certain very small distortions which are properly placed with respect to the syntactic structure or the semantic content could cause complete loss of intelligibility, while other more extensive distortions might not even be perceivable. Hence, it can be argued that objective fidelity measures which do not use semantic, syntactic, and other language related information cannot correctly predict the quality of a speech coding system.

However, an important point concerning modern speech coding systems is that, in general, they do not produce distortions which are in any way synchronous with the semantic or syntactic content of the utterance. Hence, the distortions introduced by speech coding systems represent a subset of all possible distortions. It is our hypothesis that it is possible to design relatively compact objective measures which correlate well with subjective results over this subset of distortions introduced by speech coding systems. We recognize that these measures cannot be completely general since they do not reflect the complexities of the speech perception processing.

1.3 An Approach to Designing and Testing Objective Quality Measures

Over the years, there have been numerous objective measures suggested and used for the evaluation of speech coding systems. These measures include signal-to-noise ratios, arithmetic and geometric spectral distance measures, cepstral distance measures, various parametric distance measures, such as pseudo area functions and log area functions from LPC analysis and many more.

The task of comparing and contrasting the validity of such measures is immense. To check the validity of a particular candidate objective measure over a wide class of distortions, a researcher must create a data base of distorted speech and a corresponding data base of subjective results. This is a time-consuming and expensive process, and, as a result, the validity of most commonly used objective measures remains a subject for speculation.

In general, we were interested in designing a method for comparing the validity of objective quality measures in a cost effective way. In

short, we have designed a system for measuring the quality of objective fidelity measure--i.e. a quality measure for quality measures.

The essential features of our method are illustrated in Figure 1.3-1. First, a test set of undistorted sentences is created. This set, in general, consists of phonemically balanced sentences spoken by four or more speakers. For analysis purposes, the sentences are divided into "frames" of a length of 10-30 msec. This sentence/frame set is called U(m,n), where m is the "condition" (sentence and speaker) and n is the frame number. An ensemble of distorted and coded sentences is then produced by passing the undistorted test set through a large number of controlled distortions and speech coding systems. This forms the distorted data base, D(m,n,d) (where d is the distortion) on which the objective measures will be tested.

Once the distorted data base exists, all these sentences are tested using subjective speech quality tests. These results form a data base of subjective results called S(d). A particular candidate objective measure is tested using these three data bases as follows. First, the objective quality measure is applied to all the sentences in the distorted data base. The application of the objective measure generally involves both the undistorted and distorted data bases. Then a statistical correlation analysis is done between the results of the objective measure and the subjective data base. The results of this correlation analysis are used as a figure of merit for comparing the various objective measures.

Several points should be made about this procedure. First, note that the subjective tests are only administered once regardless of how many objective measures are to be studied. Hence, the most expensive portion of this process, namely the application of the subjective tests, need only be

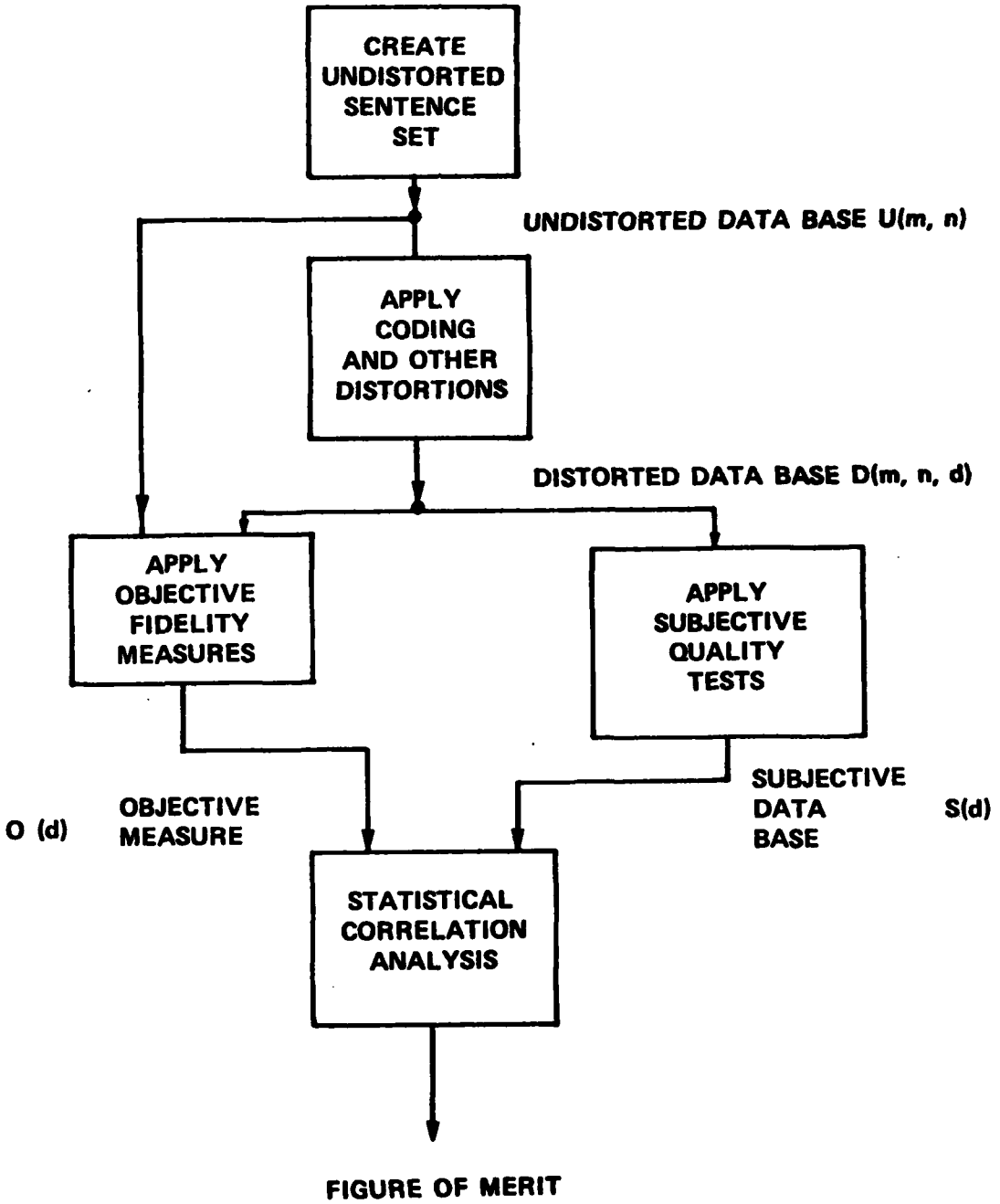


Figure 1.3-1. Block Diagram for System for Comparing the Effectiveness of Objective Quality Measures.

done once. Note also that the subjective data base may be expanded over a period of time to improve its resolving power or to extend the class of distortions involved. Similarly, subsets of the entire data base may be used if appropriate to the hypothesis being tested.

Second, note that this "quality test for quality test" system may be used to optimize the parameters of particular objective measures. This may sometimes be accomplished explicitly using statistical optimization techniques, or may be accomplished iteratively by reapplying the test repeatedly to parametrically different versions of the same objective measure.

Two figures of merit are used for a particular objective fidelity measure. The first is an estimate of the correlation coefficient between the objective fidelity measures and O(d), the subjective quality measures, S(d), given by

$$\hat{\rho} = \frac{\sum_{d} (S(d)-\overline{S(d)})(O(d)-\overline{O(d)})}{[\sum_{d} (S(d)-\overline{S(d)})^2]^{1/2}[\sum_{d} (O(d)-\overline{O(d)})^2]^{1/2}} \qquad 1.3\text{-}1$$

This results in a minimum variance linear estimate of the subjective results from the objective results given by

$$S(d) = \overline{S(d)} + \frac{\hat{\rho}\hat{\sigma}_s}{\hat{\sigma}_o} (O(d)-\overline{O(d)}) \qquad 1.3\text{-}2$$

where $\hat{\sigma}_s$ and $\hat{\sigma}_o$ are the estimated standard deviation of the subjective and objective measures, respectively. To say that this correlation

coefficient has any absolute validity would be incorrect. Since we have not randomly sampled a universe of coding distortions, our estimate of the correlation coefficient is biased. In short, estimates of correlation coefficients computed in this way are only meaningful when comparing objective measures over the same data base, and such estimates should not be compared when estimated from different data bases.

A more pleasing way to view this analysis is to view the estimate of the subjective measure as a linear regression analysis or as simply a least squares linear fit. From this, the standard deviation of the error expected when the objective estimate is used in place of the subjective estimate can be estimated by

$$\hat{\sigma}_e^2 = E[(S-E(S|O))^2] = \hat{\sigma}_s^2(1 - \hat{\rho}^2) \qquad 1.3\text{-}3$$

This estimate, which incorporates variation in the observed subjective qualities as well as the correlation coefficient, is a more pleasing figure of merit.

1.4 Principal Goals and Procedures

The research work reported here had these principal objectives:

1. To design ~1000 simple objective measures and to test their utility using correlation analysis.

2. To design both time domain and frequency domain frequency variant objective measures and to test their utility using correlation analysis.

3. To design more complex composite objective measures and to test their utility using correlation analysis.

The accomplishment of these goals involved numerous additional tasks which often led to interesting results in their own right. Some of these tasks included:

1. The design and implementation of a large data base of distorted and coded speech.

2. The performance of the subjective quality tests on the distorted data base.

3. The analysis of the subjective results directly from the distorted data base.

4. The implementation of the objective measures across the distorted and coded speech in a cost effective way.

5. The implementation of the "bulk" correlation analysis procedures necessary to handle the multitude of data produced by this effort.

In all, a total of approximately 1000 variations of simple and frequency variant measures were implemented as part of this study. These measures included simple spectral distance measures, frequency variant spectral distance measures, parametric distance measures, noise measurements, short time noise measurements, and frequency variant noise measurements. Table 1.4-1 gives a summary of the objective measures studied.

The composite objective measures considered in this study were formed by multiregression optimization on sets of the simple measures. These "complex" measures often performed much better than the simple measures, and their performance represents an estimate of the limit of the ability of objective measures to predict the results of subjective tests.

The subjective quality test used in this study was the Diagnostic Acceptability Measure (DAM) developed at the Dynastat Corporation. This test has the special feature that it provides <u>parametric</u> subjective results as well as <u>isometric</u> subjective results. This means that the objective measures may be tested as to their ability to predict these

OBJECTIVE MEASURES

SIMPLE MEASURES

SNR	6
Short Time SNR	6
Spectral Distance	192
Parametric	
Energy Ratio (Itakura)	64
PARCOR Coefficients	24
Area Ratios	24
Feedback	24
	240

FREQUENCY VARIANT

Banded SNR	6
Short Time Banded SNR	40
Spectral Distance	192
	238

COMPOSITE MEASURES	22
TOTAL	500

+Non-linear Regression	1,000
xParametric Subjective Qualities	40,000

Table 1.4-1. SUMMARY OF THE OBJECTIVE QUALITY MEASURES STUDIED

parametric results as well as the isometric results. In particular, many of the objective measures studied, including all of the frequency variant measures and the composite measures, may be "tuned" in order to predict specific parametric results. Such specific predictions, of course, are of great utility to the systems designer.

The distorted and coded speech data base consisted of 264 "distortions" which were applied to twelve sentences from each of four talkers. The total amount of speech data in these tests totaled about eighteen hours. The distortions included nine coding distortions, including both vocoder and waveform coder techniques, and fourteen "controlled" distortions, including filtering, additive noise, clipping, center clipping, interruption, echo, and frequency variant distortions. The coded distortions included both error free and fixed error rate channel simulations.

The implementation of the distorted data base, the measurement of the objective meaures, and the correlation analysis were performed on the Minicomputer Based Digital Signal Processing Laboratory [1.4] at the Georgia Institute of Technology. The subjective data base and the associated statistical analysis were performed at the Dynastat Corporation.

1.5 Summary of Major Results

One of the major characteristics of this study was that the large number of objective measures which were studied coupled with the multiple analysis methods and both the isometric and parametric subjective measures resulted in a very large number of individual correlation results (~120,000). From this large base of results, a number of specific questions were asked and answered, and a number of important results were obtained. This section will just list summaries of some of the major results.

1. A very good objective quality measure for waveform coders and noise distortions was developed based on frequency variant (banded) short time signal-to-noise measurements. This measure resulted in a correlation coefficient of .93 across all relevant distortions and a $\hat{\sigma}_e$ of 3.2 quality points on a 100 point scale.

2. The best composite measure involved some preclassification of the candidate system (vocoder vs. waveform coder), and resulted in an estimate correlation coefficient of .90 and a $\hat{\sigma}_e = 3.5$.

3. The best composite measure study which did not require preclassification had an estimated correlation coefficient of .86 and a $\hat{\sigma}_e = 4.2$.

4. Neither of the two composite measures above used higher order regression models. If such models are used, these results are improved, but there are some questions as to the accuracy of such predictions.

5. The optimum value for P in the L_P norm for spectral distance measures was found to be 8. This is a considerable departure from current practice.

6. Energy weighting of the time frame was found to have little value for any of the measures.

7. The best simple measure was found to be a log area ratio measure, which had a $\hat{\rho} = .64$ and $\hat{\sigma}_e = 6.8$. Surprisingly, this measure was better than any of the simple spectral distance measures.

8. The only two parametric measures which did well were the log area ratio measure and the energy ratio measure.

9. The frequency variant spectral distance measures performed with about a .1 point improvement in correlation over the simple measures. This was less than hoped.

10. The reliability of virtually all of the better objective measures was quite high for the number of frames used ($\sim$.99). The reliability of the subjective measures was $\sim$.9.

11. The use of higher order regression analysis (3rd order and 6th order) often gave considerable improvement in the predicted performance of the objective measures. These results, however, must be approached with caution, since some tracking of the noise is bound to be occurring.

1.6 Discussion

There are a great many aspects to this study. On the one hand, it gives, often for the first time, quantitative comparisons between many of the commonly used objective quality measures. Similarly, it gives quantitative predictions for the performance of such measures when used as predictors of subjective acceptability, at least as it is defined by the DAM test. In addition, it allows the comparison of parametrically different objective measures of the same type, and the "tuning" of individual objective measures to predict parametric subjective results. All of these results are of importance to the system's designer and the speech researcher, but, in general, do not bare directly on the overall problem of system quality measure. This is because the performance of any one measure by itself (with the noteworthy exception of the banded short time signal-to-noise ratio for waveform coders) is not good enough to effectively predict system acceptability.

On the other hand, the results of this study tell us a good deal about the performance of the subjective measures themselves, and offer new data from which to improve the subjective measures. The subjective results, in turn, can be used to judge the design of the distorted data base. These developments, once again, are quite important, but do not appreciably improve the overall quality testing environment.

The real potential for improvement comes from the use of the composite objective measures. As previously stated, this study gives fairly safe predictions of $\hat{\rho} = .86$ and $\hat{\sigma}_e = 4.2$ for such measures. There are several issues which need to be discussed here, however. First, the approach used in this study, which was necessitated by the mass of data involved, was essentially a "bulk" approach in which only standard multiregression

analysis and coarse, non-data-dependent preclassification was used. If a final "best" measure were to be designed, the results of this study should be used as a base to study the detailed behavior of the composite measures as a function of the particular distortions. Only after this is done can pragmatic variations of the composite measures be designed which allow for the special interaction of the measures with the data. Second, it should be noted that this "best" result was obtained by setting a number of parameters in the composite objective measure to optimize this measures across the distorted data base. Thus, this should be considered a limit on expected performance.

Another point concerns the nonlinear regression analysis. The number of degrees of freedom in this analysis was (usually) 1056. Hence, using 3rd order or 6th order nonlinear regression analysis was a long way from having the order of the analysis equal to the number of degrees of freedom. It is noteworthy that often remarkable improvements were obtained using nonlinear regression. Some of this effect must be noise, but clearly, some of it must be real improvement. Exactly how much improvement can be really obtained by nonlinear regression is a subject for further study.

A major point which should be made concerns the reliability of the objective measures. For the number of frames used in this study, the measured reliability was of the order of .98 or .99 for most "good" measures. This means that whatever an objective measure really measures for a distortion, it measures the same thing every time. This means that these measures could be utilized with great effectiveness for detecting malfunctions or nonstandard operation of systems in the field.

Some retrospective comment on the contents of the distorted data base is also appropriate. The data base was designed to include numerous frequency variant controlled distortions in order to facilitate the design of frequency variant objective measures. This worked well for time domain measures, but not nearly so well for frequency domain measures. Had this result been known at the outset, relatively more coding distortions would have been included.

The utility of the measures designed in this study are a function of the task for which they are to be used. This study seeks only to quantify the predicted effectiveness of objective quality measures. Thus, to determine their specific utility, one must also decide what constitutes an acceptable prediction of user acceptance.

A final point should be made here about further possible work in this area. The same techniques developed here might also be used to predict other features from subjective testing. The two most obvious classes of such tests are the parametric intelligibility tests, such as DRT, or a talker identification features test.

REFERENCES

1.1. T. P. Barnwell III, A. M. Bush, R. W. Mersereau, and R. W. Schafer, "Speech Quality Measurement," Final Report, DCA Contract No. RADC-TR-78-122, June 1977.

1.2. W. D. Voiers et al., "Methods of Predicting User Acceptance of Voice Communication Systems," Final Report, DCA 100-74-C-0056, DCA, DCEC, Reston, VA, July 1976.

1.3. W. D. Voiers, "Diagnostic Acceptability Measure for Speech Communication Systems," Conference Record, IEEE International Conference on Acoustics, Speech and Signal Processing, Hartford, CN, May 1977.

1.4. T. P. Barnwell and A. M. Bush, "A Minicomputer Based Digital Signal Processing System," EASCON '74, Washington, DC, October 1974.

CHAPTER 2

SUBJECTIVE CRITERIA OF SPEECH ACCEPTABILITY

2.1 Background

It is generally acknowledged that user acceptance of voice communications equipment depends on factors other than speech intelligibility. Intelligibility is unquestionably a necessary condition, but clearly not a sufficient condition of acceptability. Until recently, however, no generally satisfactory method of evaluating the overall acceptability of "quality" of processed or transmitted speech has been available.

Under contract with the Defense Communications Agency, Dynastat recently undertook to remedy the situation that existed in the area of acceptability evaluation. The results of this effort included the Paired Acceptability Rating Method (PARM) and the Quality Acceptance Rating Test (QUART). Both of these methods provide improved reliability of measurement on an absolute scale of acceptability, though each has limitations with respect to range of application. Both served as valuable research tools to clarify a number of crucial methodological issues and to indicate possible means of further refining the technology of speech evaluation[2.1]. Drawing on insights gained from research with these methods, Dynastat continued, under its own auspices, to further develop the technology of acceptability evaluation. These efforts have culminated with the development of the Diagnostic Acceptability Measure.

2.2 Design of the Diagnostic Acceptability Measure (DAM)

In common with several previous methods of evaluating acceptability, the DAM requires the listener to characterize transmitted speech by means of absolute, rather than relative, rating or judgments. However,

two important features distinguish it from previous methods of predicting speech acceptability. First is the fact that it combines an indirect or parametric approach with the more conventional direct or isometric approach.

In the case of the isometric approach, the listener is required to provide a simple, direct, subjective assessment of the acceptability of a sample speech transmission, for example, simply to rate a sample transmission on a 100-point scale of acceptability. Although the isometric approach has considerable appeal from the standpoint of face validity, it has several disadvantages[2.2]. For one thing, listener ratings are subject to enormous interindividual and intraindividual variation in subjective origin and scale, whether as a result of adaption level differences or simply of differences in understanding of the task. Research with PARM has shown that much of the seemingly random component of variation in rating scale data actually stems from stable listener differences in rating scale behavior. The practical implication of this finding is that differences between individual listeners or crews can seriously complicate the task. For another thing, listeners' ratings of acceptability tend strongly to be colored by differences in aesthetic preference or taste. The first of these disadvantages can be overcome to some extent through careful instructional and training procedures and by the discrete use of "anchors" and "probes." The most direct means of overcoming the second advantage is to use relatively large, representative listening crews. However, once the nature or dimensions of the interindividual differences in taste are known, stratified sampling may permit the use of smaller crews.

In the case of the parametric approach, the listener is required to evaluate the sample transmission with respect to various perceived characteristics or qualities (e.g., hissiness), ideally without regard for his personal affective reactions to these qualitities. Hence, the parametric approach serves to reduce the sampling error associated with individual differences in "tastes." An individual who does not personally place a high valuation on a particular speech quality may nevertheless provide information of use in predicting the typical individual's acceptance of speech characterized by a given degree of that quality.

A second distinguishing feature of DAM is that it solicits separate reactions from the listener with regard to what he perceives to be the speech signal itself, what he perceives to be the background, and with regard to his evaluation of the overall effect. This serves at once to reduce the listener's uncertainty as to the nature of his task and to provide the experimenter with more precise information as to the deficiencies of the system being tested. The results of many studies of human information processing indicate that, in concentrating successively on different aspects of a complex stimulus configuration, individuals are able to assimilate a greater amount of information from the stimulus--and thus respond more consistently--than otherwise.

The first step in the development of the DAM involved a series of exploratory studies designed to identify the major perceptual correlates of overall acceptability--the perceived qualities which govern the listener's acceptance reaction--and to develop the most appropriate descriptors for these correlates. This involved the experimental evaluation of a large pool of potential descriptors (e.g., hissiness) and the selection of those candidates which collectively provided the most

comprehensive and reliable discrination among various forms and degrees of speech impoverishment.

Factor analytic techniques were applied to rating data obtained with the most promising descriptors to determine the most appropriate combination of descriptors and, ultimately, to determine the nature and number of elementary perceptual qualities collectively tapped by these descriptors. Combinations of redundant descriptors were then combined to define a relatively limited number of highly discriminative rating scales. Factor analysis was used again on several occasions to further clarify the nature and number of underlying perceptual qualities and to select the combination of multidescriptor rating scales that would provide the purest and most precise measurement of each quality.

The results of several studies showed that virtually all of the perceived differences among a diversity of transmission systems and conditions could be accounted for in terms of six underlying perceptual qualities of the signal and four perceptual qualities of the background. These ten perceptual qualities were in turn found sufficient for predicting virtually all of the variation in listener ratings of the intelligibility, pleasantness, and overall acceptability of transmitted speech. It was further found that acceptability could be predicted with a high degree of precision from ratings of the two higher order qualities, perceived intelligibility and pleasantness.

The rating form shown in Figure 2.2 was developed on the basis of results of the above investigations.[1] All items on the form involve 100-

[1]Based in part on the results of the present investigation, this form will undergo several modifications for purposes of future research and services with the DAM.

DAM SYSTEM RATING FORM

Make a slash at the appropriate point on each scale to indicate the degree to which this transmission sample is characterized by the indicated quality.

THE SPEECH SIGNAL

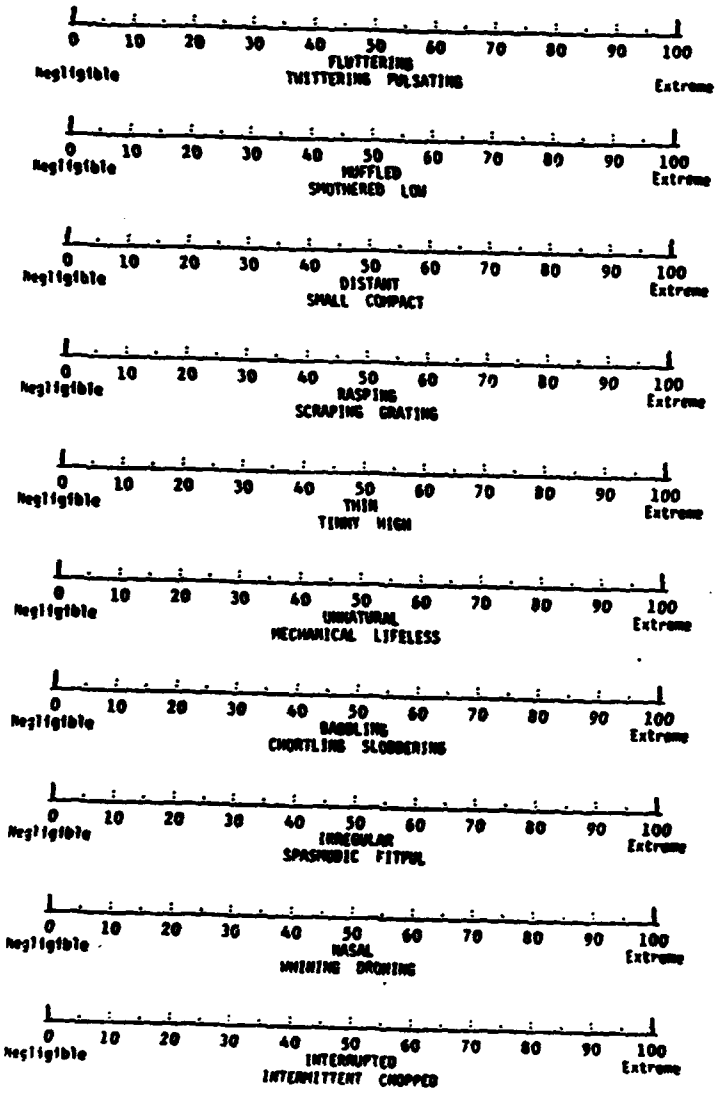
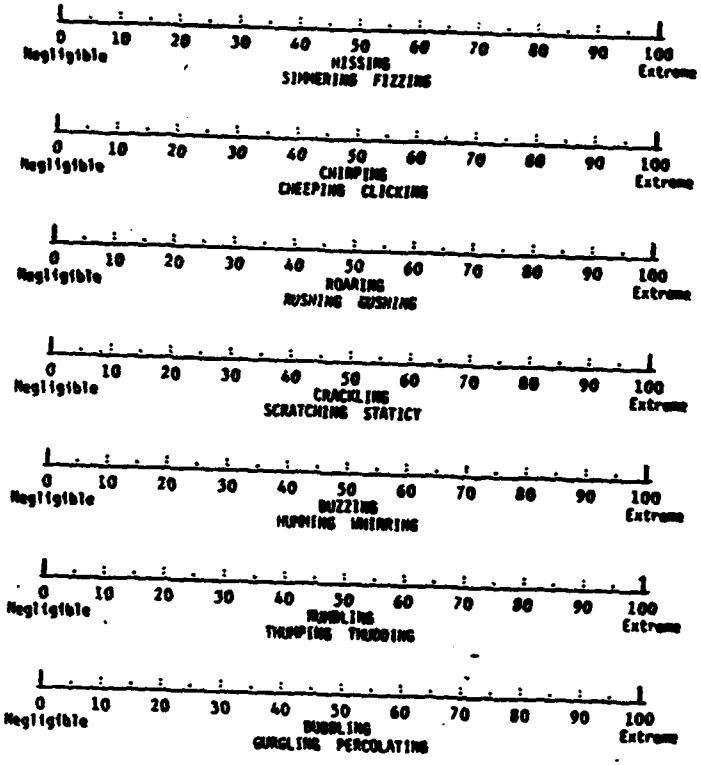


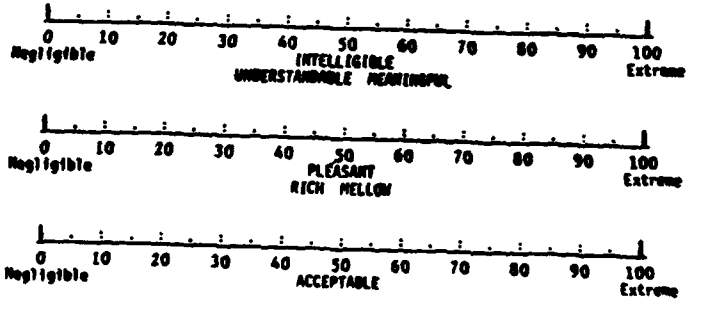
Figure 2.2 DAM Rating Form

DAM RATING FORM (cont.)

THE BACKGROUND



THE TOTAL EFFECT



point rating scales, though it should be noted that the polarities of the items pertaining to the perceptual qualities of the signal and background are the reverse of those used to evaluate overall effect. One reason for this is that most, if not all of these generally undesirable qualities, are assumed to have "true psychological zeroes." This generally is not warranted for such complex qualities as perceived pleasantness and intelligibility and overall acceptability.

Some amount of redundancy in the rating form should be evident even on casual examination. This is not an undesirable feature at this stage in the development of our knowledge of the perceptual consequences of digital voice coding. Also evident, perhaps, are the results of some attempt to provide for the perceptual consequences of yet-to-be encountered forms of speech degradation or processing. It is a reasonable expectation that features of the rating form which are redundant or extraneous at this time may find unique applicability with further developments in speech coding technology.

It follows from the above description of the rating form that more refined scoring algorithms can be developed as the need arises. For example, two of the background-rating scales clearly pertain to noise, though one would pertain most directly to high frequency noise while the other would appear to denote perceptual qualities associated with low frequency noise. For the present, these scales are combined to yield a single score for perceived background noise.

The ten perceptual qualities treated by the DAM are shown in Table 2.2-1. Each of these scoring dimensions or scales is identified by a mnemonically useful code, e.g., SL denotes that signal quality which is most conspicuously associated with "lowpassed" speech. (It should be

Table 2.2-1. STRUCTURE OF THE DAM*

Signal Quality Measures

Perceptual Quality	Rating Scales Used	Representative Descriptors	Exemplars
SF	1,7	Fluttering Bubbling	Amplitude-Modulated Speech
SH	3,5	Distant Thin	Highpassed Speech
SD	4,14	Rasping Crackling	Peak Clipped Speech, Quantized Speech
SL	2	Muffled Smothered	Lowpassed Speech
SI	8,10	Irregular Interrupted	Interrupted Speech
SN	9	Nasal Whining	Bandpassed Speech Vocoded Speech

Background Quality Measures

Perceptual Quality	Rating Scales Used	Representative Descriptors	Exemplars
BN	11,13	Hissing Rushing	Guassian Noise
BB	15	Buzzing Humming	60-120 Hz Hum
BF	12,17	Chirping Bubbling	Errors in narrow band systems
BR	16	Rumbling Thumping	Low frequency noise

Total Quality Measures

Quality	Rating Scales Used	Representative Descriptors	Exemplars
Intelligibility	18	Intelligible	Undegraded Speech
Pleasantness	19	Pleasant	Undegraded Speech
Acceptability	20	Acceptable	Undegraded Speech

stressed, however, that lowpassing of speech has perceptual consequences other than those reflected on the SL scale, and, moreover, that SL scores may be affected by other conditions than high frequency attenuation.)

For greater convenience in interpretation of score patterns, the polarities of the ten derived scales are reversed from those of the original seventeen rating scales. High scores on the derived scales are thus associated with freedom from the various perceptual qualities; and are thus associated with acceptability, as is the case with ratings of intelligibility, pleasantness and acceptability, itself.

The contribution of each perceptual quality to the listener's acceptance reaction has been closely approximated through experimentation, so that each diagnostic score represents the estimated level of acceptability a system would be accorded if it were deficient with respect only to the single perceptual quality involved. Thus, the pattern of diagnostic scores provides estimates of the relative contributions of the ten perceptual qualities to the acceptance of the system, and permits the communications engineer to identify the characteristics of a system or device which are most detrimental to its acceptance, regardless of difference in the values listeners place on the various qualities.

The application of a multiple nonlinear regression equation (based on an analysis of DAM data for more than 200 system-conditions) to the ten diagnostic scores yields one gross parametric estimate of the acceptability of the system or condition being evaluated. Appropriately transformed ratings of intelligibility and pleasantness provide two additional parametric estimates. (These transformations take into account the fact that acceptability is a slightly positively accelerated function of judged intelligibility while being a negatively increasing function of judged

pleasantness.) The three parametric estimates are then averaged with raw or isometric ratings of acceptability to provide the one best, composite estimate of acceptability.

To permit comparisons with the results of previous evaluations obtained with PARM, composite acceptability estimates are transformed to their PARM equivalents on the basis of the observed regression of PARM scores on DAM composite scores in a sample of more than 200 system conditions. A relatively crude estimate of intelligibility is obtained from intelligibility ratings based on the regression of DRT scores on these ratings in a sample of approximately 100 system conditions (actual speech coding systems.)

2.3 Materials and Procedures

2.3.1 Speech Materials

The test speech material used with the DAM consisted of twelve phonemically controlled six-syllable sentences [2.1] which are uttered by speakers at a rate of one sentence per four seconds. Different sentences are used by different speakers, but the same twelve sentences are always spoken by each speaker.

2.3.2 Evaluation Procedures

From six to twenty-four experimental system-conditions may be evaluated in the course of one testing session, depending on the number of speakers involved. Ideally, listeners evaluate all system-conditions in sub-sessions involving one speaker at a time. It is particularly desirable, however, that the time-ordering of the conditions varies from one speaker to the next in a counter-balanced manner. At the beginning of each sub-session, listeners evaluate two "anchors" and four "probes." The

purpose of the anchors is to provide the listeners with a frame of reference in which to make their ratings of experimental system conditions. Data from the four probes, (an LPC, a CVSD, a channel vocoder, and Parkhill) are used to adjust all rating data for any circumstantial factors which may have operated to increase or decrease the average of all system ratings for a given sub-session. Where average ratings of the four probes on any scale deviate from historical norms, all data for that scale are adjusted in the opposite direction. But, due to the fact that deviations in averaged probe ratings do not provide perfectly reliable measures of changes in the crews subjective origin or adaptation level, ratings of system conditions and the probes themselves are adjusted by an amount equal to only .5 of the probe deviation from historical norms.

2.3.3 Listener selection and calibration

Listeners used for system evaluations with the DAM undergo rigorous selection and training procedures. Initial selection is achieved with the use of the DAM itself. Candidates make ratings of a diversity of system conditions. The correlations between the candidated ratings and normative ratings provide the basis of selection. Following learning sessions with a diversity of system-conditions, listener trainees undergo a calibration session in which they rate a highly diverse sample of more than 200 system-conditions with three speakers for each condition.

The regressions of individual listener ratings on normative rating values provide the basis for adjusting the individual's data to compensate for differences between his subjective origins and scales and those of the historical normative listener. Coefficients of correlation obtained in the course of this analysis determine the relative weight accorded the individual listener's data in subsequent tests and experiments. Listeners

are periodically recalibrated to adjust for changes in their response characteristics that may occur with time and experience.

2.3.4 Analysis of DAM data

The first step in the analysis of DAM data involves the inversion of signal and background quality rating data for each listener.

$$R'_{ij(u)} = 90 - R_{ij} \qquad 2.3.4\text{-}1$$

where $R'_{ij(u)}$ is an inverted rating datum for the jth condition on the ith rating scale, 90 is the historically-normative inverted rating of the high anchor on the ith rating scale and R_{ij} is a raw rating of the jth condition on the ith scale. All values are further transformed such that:

$$R''_{ij(u)} = b_i R_{ij} + C_i \qquad 2.3.4\text{-}2$$

where b_i and C_i are selected such that $R_{ij(u)}$ closely approximates the acceptability rating condition j would receive it its sole deficiency were in terms of the system characteristic tapped by scale i. Values for R_i for various scales are then used singly or averaged in various combinations to yield unadjusted (for listener idiosyncracies) perceptual quality values, (S_{ij} for each listener condition).

Values of $S_{ij(u)}$ for each listener, k, are transformed as follows:

$$S'_{ijk} = b_{ik} S_{ijk} + C_{ik} \qquad 2.3.4\text{-}3$$

where b_{ik} is a scale factor which relates listener k to the normative listener for perceptual quality scale, i, and C_{ik} is the difference in

subjective origin between listener k and the normative listener. A weighted average:

$$\frac{\sum_{k=1}^{n} \left(r_{ik} \, S'_{ijk} \right)}{\sum_{k=1}^{n} r_{ik}} \qquad 2.3.4\text{-}4$$

where r_{ik} is the correlation between listener k's rating on a scale i of a standard set of conditions and the historically normative ratings of the same set of conditions.[2] The effect of this process is to give greatest weight to those listeners whose response characteristics correlate most highly with those of the historically normative listener.

A final, minor adjustment of all averaged adjusted perceptual quality values is made in an effort to control transient circumstantial influences to which the crew as a whole may be subject during a given experimental session. This is accomplished by means of the formula:

$$\bar{S}_{ij(p)} = \bar{S}_{ij} - .5 \, (\bar{P}_i - \bar{P}_{i(h)}) \qquad 2.3.4\text{-}5$$

where $\bar{S}_{ij(p)}$ is the "probe-adjusted" crew average rating of condition j on perceptual quality, i, $\bar{P}_i$ is the presently obtained average rating of the four probes and $\bar{P}_{i(h)}$ is the historical average rating of the same crew's

---

[1]The bar over the subscript $\bar{i}$ is used here to indicate that perceptual quality scale values are in some instances obtained by averaging two transformed rating scale values. Henceforth, i will be used without the bar to denote the perceptual qualitites, themselves, rather than the rating scales from which estimates of them are obtained.

[2]The normal symbological convention in statistics is that the subscripts to r_{ik} denote the two correlated variables. This convention is not observed in this instance alone.

ratings of the four probes, and .5 is the estimated coefficient of reliability (session to session) of the probe average. Fully adjusted perceptual quality averages serve, as such, for purposes of detailed system diagnosis, but they also provide the basis for estimates of three higher-order criteria of system performance: total signal quality (TSQ), total background quality (TBQ) and a parametric estimate of overall system acceptability (PA). These measures are derived by means of the following equations:

$$TSQ = C_i\left[\left(\sum_{i=1}^{b} b_i S_i + C_3 \prod_{i=1}^{6} i/10\right) - C_3\right]$$

$$TBQ = C_i\left[\left(\sum_{i=7}^{10} b_i S_i + C_3 \prod_{i=7}^{10}\right) - C_3\right] \qquad 2.3.4\text{-}6$$

(Corresponding constants in the two equations are not identical, but C_i is in each case designed to transform the measure in question into its acceptability equivalent e.g., the acceptability level the system would be accorded if its deficiencies were confined to perceived signal qualities.)

$$PA = \sum_{i=1}^{10} b_i S_i + C_1 (TSQ \times TBQ) + C_2 \qquad 2.3.4\text{-}7$$

where the regression coefficients regression constants have been estimated on the basis of data for more than 200 system conditions. Even with a sample of this size, however, it is to be expected that minor adjustments of the b_i's and constants, and of the form of these equations will be made as more DAM data are accumulated.

Two additional parametric estimates of acceptability are derived from isometric ratings of intelligibility and pleasantness.

$$PI = C_1 I + C_2 I^2 + C_3 \qquad 2.3.4\text{-}8$$

$$PP = C_1 P + C_2 P^2 + C_3 \qquad 2.3.4\text{-}9$$

where I and P are averaged ratings of intelligibility and pleasantness which have been adjusted for listener idiosyncracies and circumstantial effects in the same manner as the perceptual quality values.

Direct, isometric, ratings of acceptability provide the last of the four gross estimates of system acceptability. Following adjustments for listener idiosyncracies, the isometric estimate of system acceptability is averaged with PA, PI, and PP to obtain the best composite estimate, CA, of overall acceptability. Due to slight differences in the reliabilities of these four estimates--PA has a slightly higher reliability (.976) than the other three measures--a weighted averaged is used for this purpose.

REFERENCES

2.1 W. D. Voiers and staff of Dynastat, Inc., "Methods of Predicting User Acceptance of Voice Communication Systems," Final Report, DCA 100-74-C-0056, DCA, DCEC, Reston, VA.

2.2 W. D. Voiers, "Diagnostic Acceptablity Measure for Speech Communications Systems," Conference Record, IEEE ICASSP, Hartford, CN, May 1977.

CHAPTER 3

OBJECTIVE MEASURES

3.1 Introduction

Three of the goals of this study as discussed in Chapter 1 were: (1) to identify a set of promising objective measures for speech quality; (2) to test these measures in order to quantity their effectiveness as speech fidelity measures; and (3) to design new measures which are better able to predict the results of subjective speech quality measures. The purpose of this chapter is to describe in detail the "basic" objective measures considered in this study.

In the past several years, there has been considerable interest in defining and using objective measures for speech quality [3.1]. As was discussed in Chapter 1, the two main uses of objective quality measures are the prediction of user acceptance of candidate coding systems and the "optimization" of coding systems using the objective quality measures as fidelity criteria. The first use leads to reduction in cost of subjective quality testing, while the second leads to higher quality speech communications systems.

The objective measures included in this study were mainly intended for the testing of the three main classes of digital coding systems: waveform coders, in which the coding system tries to duplicate the input signal at the output; vocoders, in which the system does a deconvolution of the filtering effect of the upper vocal tract from the excitation function; and transform coding, where a two dimensional time-frequency representation of the speech waveform is coded instead of the waveform itself.

This bias toward digital systems is mainly motivated by current trends in technology. This does not mean that the results here are not applicable to analog systems, but such systems do pose somewhat greater problems in synchronization and phase control.

The objective measures studied here can be divided roughly into six classes: simple spectral distance; simple noise; parametric; frequency variant spectral distance; frequency variant noise; and composite. Simple spectral distance measures includes all those measures in which the distortion is computed entirely in the frequency domain and in which the spectral weighting of the measure is either unity or derived from the original speech signal. Simple noise measures include all those measures in which the main component is the "noise" between the input speech signal and the output coded signal computed entirely in the time domain. Parametric measures include all those measures in which the measure is derived from some secondary parameter set which has been derived from the speech signals under test. In frequency variant spectral distance measures, the measures are performed in the frequency domain, but are performed in bands rather than across the entire frequency range. In frequency variant noise measures the noise is measured in predetermined frequency bands by appropriate pre-filtering. Composite measures are new, hopefully improved, measures derived by combining measures from the other five classes.

The two classes of "simple" measures and the parametric measures are included for three principal reasons. First, they are to quantity the effectiveness of many of the measures currently in common use for speech quality prediction. Second, they are to test the effect of parametrically different forms of the various measures. Finally, they are to test the utility of such measures against more complex measures.

The two frequency variant classes of measures are included for two principal reasons. First, it has been known for some time [3.4] that hearing and speech perception are a frequency variant operation. This phenomenon has been studied physically, but the measurement of precise physical parameters is very difficult. The frequency variant measures form a domain in which a secondary measurement of these effects can be made using correlation analysis [3.3]. Second, it is well known that many of the parametric subjective measures from the DAM (see Chapter 2) are frequency related. The frequency variant objective measures form a domain in which the objective measures may be "tuned" to predict such parametric subjective quality results.

The design of the composite measures is one of the principal goals of this study. Composite measures are specially intended to be used in future objective-subjective testing and as diagnostic tools for coding systems.

3.2 Basic Concepts and Notations

Objective measures are made between an undistorted speech data set, ϕ, and a distorted speech data set, d. In this study, the undistorted speech data set is made up of a four speaker set, s. Each basic speech set consists of twelve sentences from each of the four speakers (see Chapter 4 for more details).

In computing objective measures, the estimate is generally formed by averaging the results from a number of "frames" of the undistorted and distorted speech. In order for the measures to be unbiased, precise frame synchronization between the distorted and undistorted speech signal must be maintained. Since all of the distortions in this study were digitally produced, synchronization was not a great problem during this study (see

Chapter 1 and Chapter 4). However for the testing of non-simulated coding systems, the synchronization problem would have to be carefully considered.

The objective measures in this study are computed from a set of input undistorted speech frames, $X(n,s,\phi)$, where n is the frame index, s is the speaker index and ϕ means no distortion, and a distorted speech set, $X(n,s,d)$, where d is the distortion. Here, the distortion may mean coding distortion or a controlled distortion (Chapter 4). In general, each distortion measure is characterized by a specific function, F at the frame level; and, in general, all the objective measures, called O(d), are computed from

$$O(d) = \frac{\sum_{s=1}^{4} \sum_{n=1}^{N} W(n,s)\, F[X(n,s,\phi),X(n,s,d)]}{\sum_{s=1}^{4} \sum_{n=1}^{N} W(n,s)} \qquad 3.2\text{-}1$$

where N is the number of frames in the analysis, and W(n,s) is a weighting function for the $n^{\underline{th}}$ frame and the $s^{\underline{th}}$ speaker. Note that W(n,s) may also be a function of $X(n,s,\phi)$, $X(n,s,d)$, or both. In this environment, therefore, describing the objective measures reduces to describing the functions W(n,s) and $F[X(n,s,\phi),X(n,s,d)]$ used for each measure.

3.3 The Simple Measures

The simple measures refer to the set of measures which produce an isometric quality measure from a single compact computational algorithm. These measures include such traditional measures as SNR, spectral distance, etc. This section describes measures of this type used in this study.

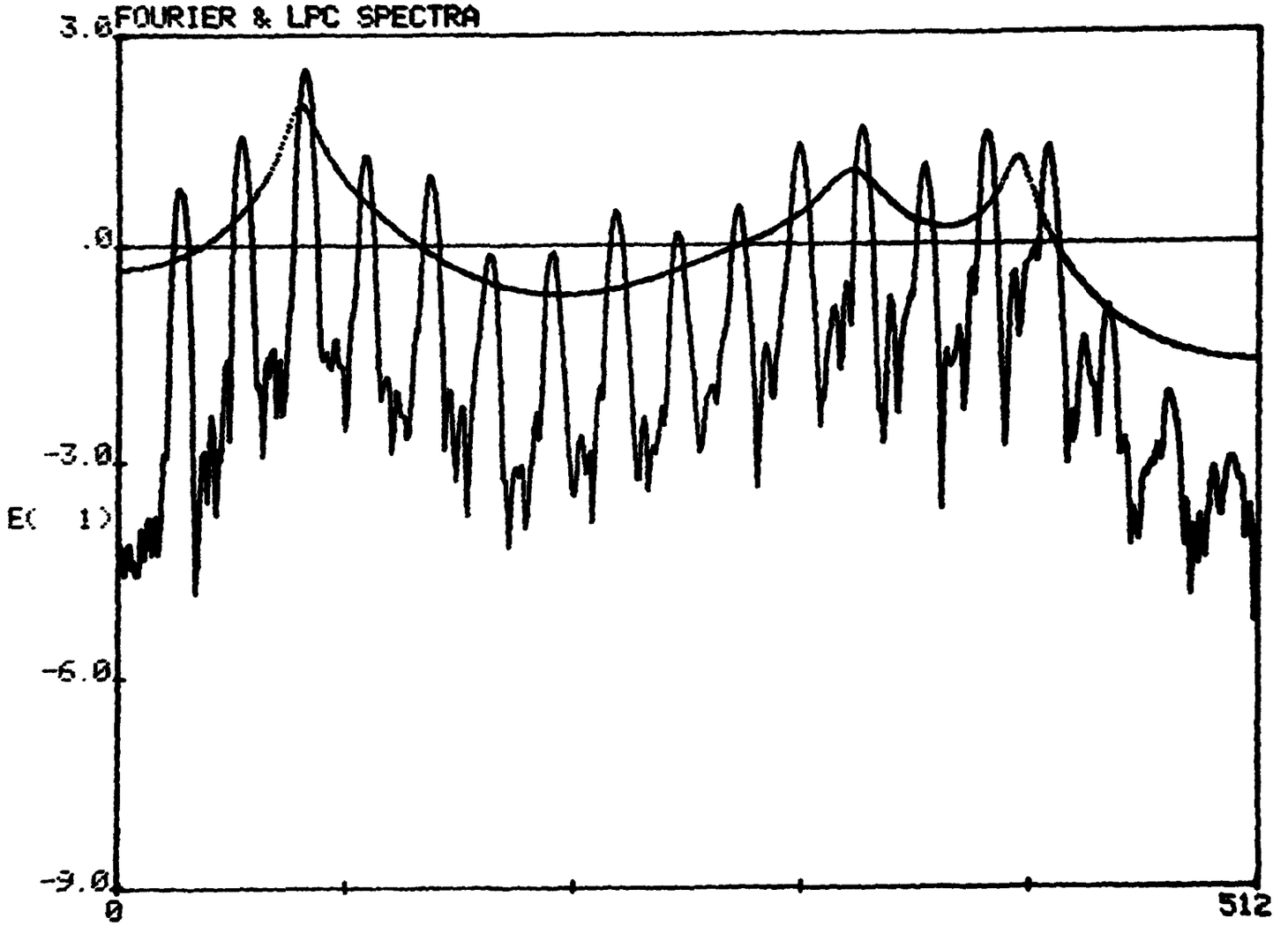


Figure 3.3.1-1. Comparison of Fourier and LPC Spectra for a Vowel.

3.3.1 The Spectral Distance Measures

All spectral distance measures are based on a function $V(n,s,d,\theta)$, the "spectrum" for the $n^{\underline{th}}$ frame speaker s, the $d^{\underline{th}}$ distortion, and the frequency variable, θ. The first question to be answered is how to derive this spectrum from the input speech sample $X(n,s,d)$. Let $x(m,s,d)$ be the sampled (at 8 kHz) digital representation of the distorted signal for the $s^{\underline{th}}$ speaker and the $d^{\underline{th}}$ distortion. Then the "framed" speech time sample for the $n^{\underline{th}}$ frame, $x_n(m,s,d)$, is given by

$$x_n(m,s,d) = x(m,s,d)\ W(m-nI) \qquad 3.3.1\text{-}1$$

where $W(m)$ is a finite length window function and I is the frame interval in samples. The Discrete Fourier Spectrum for this signal is given by

$$V(n,s,d,\theta) = \left| \sum_{m=-\infty}^{+\infty} x_n(m,s,d)e^{-j\theta m} \right| \qquad 3.3.1\text{-}2$$

where the limits on the sum are really finite because of the finite length of $x_n(m,s,d)$. The short time stationarity of speech [3.4] suggests that a good window length is 10-30 msec. Although the DFT is a very natural function to consider, there are several arguments against its use. First, for the window lengths above $x_n(m,s,d)$ would normally include several pitch periods. This would cause $V(n,s,d,\theta)$ to be a line spectrum, as shown in Figure 3.3.1-1. Because small variations in pitch, which have little impact on quality, would cause great differences between such spectra, then the DFT is not a good candidate for a spectral distance measure. What is really needed is the spectral envelope of the DFT. This can be approximated in several ways. First, it can be approximated by always having only

one pitch period in the analysis window of the DFT. This method, however, would need the use of a pitch detector plus additional synchronization logic which makes this approach unattractive. Second, the spectral envelope can be estimated using the parametric LPC analysis technique [3.5],[3.6],[3.7]. The advantage of this technique is that it is computationally simple and results in a very compact representation of the spectral envelope. However, like all parametric approaches, it is subject to modeling errors. Finally, the spectral envelope could be extracted using cepstral deconvolution techniques [3.8],[3.9] . However, previous research has shown [3.1],[3.10] that this measure is very highly correlated with the corresponding LPC technique and cepstral analysis is more computationally intense.

3.3.1.1 The LPC Parametric Analysis Technique

In this study, the basis for the spectral envelope approximations was always the LPC parametric technique. In this technique, a set of autocorrelation functions, given by

$$R_n(k) = \sum_{m=-\infty}^{+\infty} x_n(m,s,d)\, x_n(m+k,s,d) \qquad 3.3.1.1\text{-}1$$

for the $n^{\underline{th}}$ frame and $0 \le k \le 10$, are computed, and then a set of 10 "feedback coefficients," a(k), are computed from Durbin's recursion, given by

$$\begin{aligned} &\alpha(n) = R(0);\ K(1) = -R(1)/R(0);\ a^1(1) = -K(1) \\ &\alpha(n) = (1 - K^2(n-1))\ (n-1) \\ &K(n) = \sum_{i=1}^{n-1} (a^{n-1}(i)R(n-i)-R(n))/\alpha(n) \\ &a^n(n) = -K(n);\ a^n(i) = a^{n-1}(i) + K(n)a^{n-1}(n-i) \end{aligned} \qquad 3.3.1.1\text{-}2$$

where the autocorrrelation subscripts have been dropped. In this recursion, the K(n) parameters are the well-known PARCOR (partial correlation coefficients) first used by Itakura [3.11]. From the feedback coefficients, the energy spectrum can be computed by

$$V(n,s,d,\theta) = \left| \frac{G}{1 - \sum_{k=1}^{10} a(k)e^{-j\theta k}} \right| \qquad 3.3.1.1\text{-}3$$

where G is the gain term, given by

$$G = [R(0) - \sum_{k=1}^{10} a(k)R(k)]^{1/2}. \qquad 3.3.1.1\text{-}4$$

The LPC approach has several specific advantages when used for spectral analysis. First, the entire analysis for a frame results in only 11 numbers, a(1) - a(10), and G. This means that a large number of spectral analysis results may be stored relatively compactly. Second, the gain analysis is separate from the spectral analysis. Since small changes in gain do not have great impact on perception, it is desirable to remove gain effects from the spectral distance measure. One reasonable way in which this may be done from the LPC analysis is force the gain term in Equation 3.3.1.1-2 to be 1, giving

$$V(n,s,d,\theta) = \left| \frac{1}{1 - \sum_{k=1}^{10} a(k)e^{-j\theta n}} \right| \qquad 3.3.1.1\text{-}5$$

This normalizes the total area under the $V(n,s,d,\theta)$ to be equal to 1. Finally, the LPC method results in a relatively compact computation of $V(n,s,d,\theta)$ from a(1) - a(10). $V(n,s,d,\theta)$ may be thought of as the

magnitude of the discrete Fourier transform (DFT) of the impulse response of an infinite impulse response filter (IIR) whose Z transform is given by

$$V(Z) = \frac{1}{1 - \sum_{k=1}^{10} a(k)Z^{-k}} \qquad 3.3.1.1\text{-}6$$

The inverse of this filter is an FIR (finite impulse response) filter whose Z transform, I(Z), is given by

$$I(Z) = \frac{1}{V(Z)} = 1 - \sum_{k=1}^{10} a(k)Z^{-k}. \qquad 3.3.1.1\text{-}7$$

The spectrum for $I(n,s,d,\theta)$, the inverse of $V(n,s,d,\theta)$, can hence be computed from

$$I(n,s,d,\theta) = \left| 1 - \sum_{k=1}^{10} a(k)e^{-j\theta k} \right| . \qquad 3.3.1.1\text{-}8$$

Since this sum has only 11 terms, it can be computed very compactly. Even greater gains may be obtained if the FFT is used. Once $I(n,s,d,\theta)$ is known, $V(n,s,d,\theta)$ may be simply obtained from

$$V(n,s,d,\theta) = 1/I(n,s,d,\theta). \qquad 3.3.1.1\text{-}9$$

3.3.1.2 The Computation of Objective Measures

In this study, six variations of the distance function for spectral distance analysis, i.e. the function F in Equation 3.2-1, were studied. The first, called the "linear unweighted" spectral distance, is given by

$$F = \left[\frac{1}{L} \sum_{\ell=0}^{L-1} [V(n,s,\phi,\theta_\ell) - V(n,s,d,\theta_\ell)]^p \right]^{1/p} \qquad 3.3.1.2\text{-}1$$

i.e. the L_p norm of the sample difference. In general, L=128 and

$$\theta_\ell = \frac{\pi \ell}{L} \qquad = 0,\ldots,L\text{-}1 \qquad 3.3.1.2\text{-}2$$

The second form, called the "linear frequency weighted" form, is given by

$$F = \left[\frac{\sum_{\ell=0}^{L} |V(n,s,\phi,\theta_\ell)|^\gamma \, |V(n,s,\phi,\theta_\ell) - V(n,s,d,\theta_\ell)|^p}{\sum_{\ell=1}^{L} |V(n,s,\phi,\theta_\ell)|^\gamma} \right]^{1/p} \qquad 3.3.1.2\text{-}3$$

In this form, the measure is weighted by the spectrum of the undistorted spectrum taken to the γ power. The third form, called the "log unweighted" spectral distance is given by

$$F = \left[\frac{1}{L} \sum_{\ell=0}^{L-1} \left| 20 \log_{10} \left[\frac{V(n,s,\phi,\theta_\ell)}{V(n,s,d,\theta_\ell)} \right] \right|^p \right]^{1/p} \qquad 3.3.1.2\text{-}4$$

Here the constant 20 is used to produce results in db. The fourth form, the "frequency weighted log" spectral distance measure is given by

$$F = \left[\frac{\sum_{\ell=0}^{L-1} |V(n,s,\phi,\theta_\ell)|^\gamma \left| 20 \log_{10} \left[\frac{V(n,s,\phi,\theta_\ell)}{V(m.s,d,\theta_\ell)} \right] \right|^p}{\sum_{\ell=0}^{L-1} |V(n,s,\phi,\theta_\ell)|^\gamma} \right]^{1/p} \qquad 3.3.1.2\text{-}5$$

The fifth form of the spectral distance measure, called the "unweighted δ" form is given by

$$F = \left[\frac{1}{L} \sum_{\ell=0}^{L-1} |V(n,s,\phi,\theta_\ell)^\delta - V(n,s,d,\theta_\ell)^\delta|^p \right]^{1/p} \qquad 3.3.1.2\text{-}6$$

Finally, the "frequency weighted δ" form is given by

$$F = \left[\frac{\sum_{\ell=0}^{L-1} |V(n,s,\phi,\theta_\ell)|^\gamma \left| V(n,s,\phi,\theta_\ell)^\delta - V(n,s,d,\theta_\ell)^\delta \right|^p}{\sum_{\ell=0}^{L-1} |V(n,s,\phi,\theta_\ell)|^\gamma} \right]^{1/p} \qquad 3.3.1.2\text{-}7$$

Implicit in the definitions of the spectral distances above are three major questions. First, what nonlinearity should be applied to the spectrums before computing the distances for best results? The three candidates here are none (linear), log, and raising the spectrum to the δ power. This last form is an approximate bridge between the other two forms. Second, should the spectrum be weighted by a function of the undistorted spectrum, and, if so, by how much? The control parameter for this case is γ. Finally, what value of p for the L_p norm should be used? For this case, as $p \rightarrow \infty$, the criterion approaches minimax.

3.3.2 Parametric Distance Measures

As in the case of spectral distance measures, the parametric distance measures assume that the distorted and undistorted speech signal has been divided into frames, given by $X(n,s,\phi)$ and $X(n,s,d)$ where n is the frame number, s is the speaker, d is the distortion, and ϕ indicates no

distortion. For each parametric distance measure, a set of L parameters, $\xi(n,s,d,\ell)$, $\ell = 1,\ldots,L$, are derived from the corresponding speech frame X(n,s,d). As in the case of spectral distance, a function F for use in Equation 3.2-1 is derived for each case, given by

$$F = \left[\frac{1}{L} \sum_{\ell=1}^{L} |\xi(n,s,d,\phi) - \xi(n,s,d,\ell)|^p \right]^{1/p} \qquad 3.3.2\text{-}1$$

where once again the L_p norm is taken. As before, p is an object of study for each parametric distance measure.

All of the parametric distance measures studied were derivatives of LPC analysis. There were eight basic measures considered in this study. The first two were based on the feedback coefficients set, a(1)-a(10), which is described in Equation 3.3.1.1-2. The first form, the "linear feedback" measure is given by

$$F = \left[\frac{1}{10} \sum_{\ell=1}^{10} |a(n,s,d,\ell) - a(n,s,d,\phi)|^p \right]^{1/p} \qquad 3.3.2\text{-}2$$

and second form, the "log feedback" measure is given by

$$F = \left[\frac{1}{10} \sum_{\ell=1}^{10} \left| 20 \log_{10} \left| \frac{a(n,s,d,\ell)}{a(n,s,\phi,\ell)} \right| \right|^p \right]^{1/p} \qquad 3.3.2\text{-}3$$

This second measure was not expected to be of much interest, but was included for completeness.

The third and fourth measures were based on the PARCOR coefficients, K(m), as defined in Equation 3.3.1.1-3. These two measures are given by

$$F = \left[\frac{1}{10} \sum_{\ell=1}^{10} \left| K(n,s,d,\ell) - K(n,s,\phi,\ell) \right|^p \right]^{1/p} \qquad 3.3.2\text{-}4$$

and

$$F = \left[\frac{1}{10} \sum_{\ell=1}^{10} \left| 20 \log_{10} \left| \frac{K(n,s,d,\ell)}{K(n,s,\phi,\ell)} \right| \right|^p \right]^{1/p} \qquad 3.3.2\text{-}5$$

where $K(n,s,d,\ell)$ and $K(n,s,\phi,\ell)$ are the ℓth PARCOR coefficients derived from the (n,s) frame of the distorted and undistorted speech sample, respectively.

The fifth, sixth, and seventh measures were based on the area ratios functions $AR(n,s,d,\ell)$ given by

$$AR(n,s,d,\ell) = \frac{1 - K(n,s,d,\ell)}{1 + K(n,s,d,\ell)} \qquad 3.3.2\text{-}6$$

These measures are given by

$$F = \left[\frac{1}{10} \sum_{\ell=1}^{10} |AR(n,s,\phi,\ell) - AR(n,s,d,\ell)|^p \right]^{1/p} \qquad 3.3.2\text{-}7$$

and

$$F = \left[\frac{1}{10} \sum_{\ell=1}^{10} 20 \log_{10} \left| \frac{AR(n,s,d,\ell)}{AR(n,s,\phi,\ell)} \right|^p \right]^{1/p} \qquad 3.3.2\text{-}8$$

and

$$F = \frac{1}{10} \sum_{\ell=1}^{10} |AR(n,s,d,\ell)^{\delta} - AR(n,s,\phi,\ell)^{\delta}|^p \quad ^{1/p} \qquad 3.3.2\text{-}9$$

The final parametric measure of interest is called the "energy ratio" measure which was first suggested by Itakura [3.11], and has been widely used as a quality measure [3.12],[3.13]. In this analysis, a frame by frame LPC analysis is performed on both the undistorted and distorted speech, as shown. Then undistorted speech is passed through two "vocal track inverse filters" given by

$$H(Z) = 1 - \sum_{\ell=1}^{10} a(\ell) Z^{-\ell}$$

$$H'(Z) = 1 - \sum_{\ell=1}^{10} a'(\ell) Z^{-\ell} \qquad 3.3.2\text{-}10$$

The energy out of each channel is squared and summed, given $e^2(n,s,\phi)$ and $e^2(n,s,d)$. The energy ratio is then given by

$$F = \frac{e^2(n,s,d)}{e^2(n,s,\phi)} \qquad 3.3.2\text{-}11$$

or

$$F = 20 \log_{10} \frac{e(n,s,d)}{e(n,s,\phi)} \qquad 3.3.2\text{-}12$$

It turns out that this measure can be computed more compactly than is suggested by the above results. In particular, it can be shown that

$$\frac{e(n,s,d)}{e(n,s,\phi)} = \left[\frac{\underline{A}^T(n,s,d)\ \underline{\underline{R}}(n,s,\phi)\ \underline{A}(n,s,d)}{\underline{A}^T(n,s,\phi)\ \underline{\underline{R}}(n,s,\phi)\ \underline{A}(n,s,\phi)} \right]^{1/2} \qquad 3.3.2\text{-}13$$

where

$$\underline{\underline{R}} = \begin{bmatrix} R(0) & R(1) & . & . & . & R(9) \\ R(1) & R(0) & . & . & . & R(8) \\ R(2) & R(1) & R(0) & . & . & R(7) \\ . & . & . & . & . & \\ . & . & . & . & . & \\ R(9) & . & . & . & . & R(0) \end{bmatrix} \qquad 3.3.2\text{-}14$$

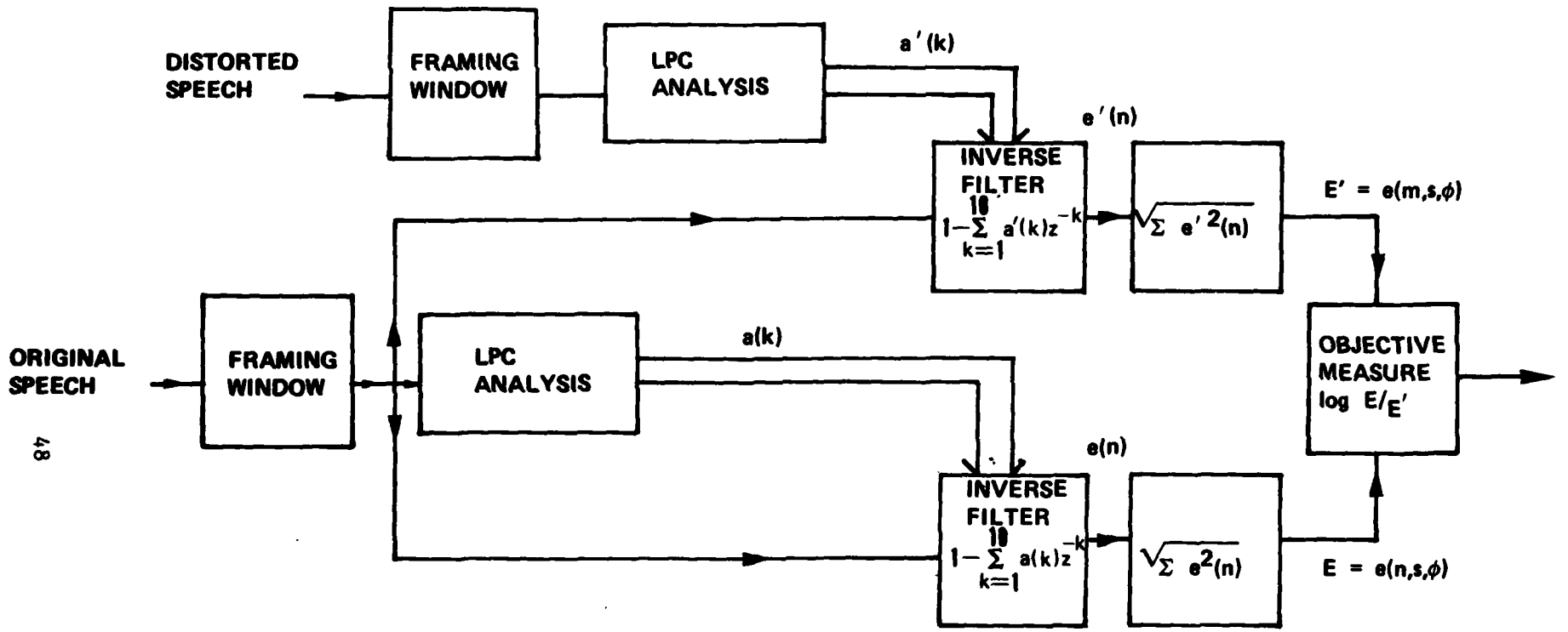


Figure 3.3.2–1 Computation of the Residual Energy Distance Measure

and R(k) is defined by Equation 3.3.1.1-1 and

$$\underline{A} = \begin{bmatrix} a(1) \\ a(2) \\ \cdot \\ \cdot \\ \cdot \\ a(10) \end{bmatrix} \qquad 3.3.2\text{-}15$$

where a(k) is defined by Equation 3.3.1.1-2. The three forms of this measure which were studied are given by Equations 3.3.2-11 and 3.3.2-12, plus

$$F = \left| \frac{e(n,s,d)}{e(n,s,\phi)} \right|^{\delta} \qquad 3.3.2\text{-}16$$

The parametric distance measure study had three main goals. First, to compare the various types of parameters for their ability to predict subjective results. Second, to investigate the value of p for the L_p norms which gives the best results. Finally, to investigate the nonlinearity (none, log, or $|.|^{\delta}$) which is most appropriate for good prediction of subjective results.

3.3.3 Simple Noise Measurements

For many years, the signal-to-noise ratio (SNR) has been used as a quality measure for systems in which it is an applicable concept. In digital communications, the signal plus noise model is meaningful in systems where the received signal is designed to be a point by point copy

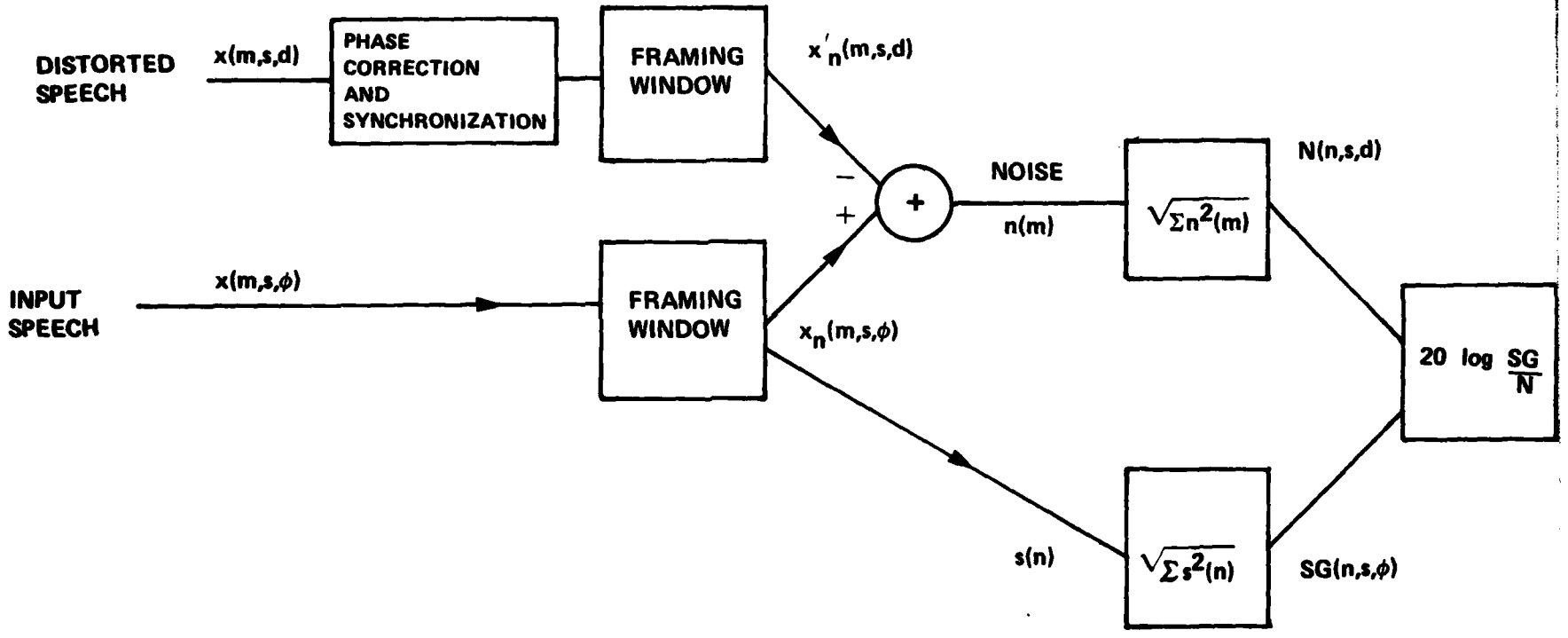


Figure 3.3.3–1 System for computing short time SNR

of the input signal. These systems include all forms of waveform coders, including CVSD, ADM, DPCM, ADPCM, and APC, as well as such new techniques as sub-band coding and adaptive transform coding. These systems do not include the vocoder and "vocoder-like" systems such as LPC, VEV's of all types, channel vocoders, etc.

In this study, two types of broadband noise measurements were studied. The first was the traditional SNR. In this system (see Fig. 3.3.3-1) any linear or nonlinear phase variations introduced in processing are first corrected. Since all of the distortions in the study were produced by computer simulation, this process was a completely tractable procedure. If real digital communications systems were to be tested, the synchronization and nonlinear phase correction problem could be very great. Once the phase corrected signals are available, the frame noise energy, N(n,s,d) is computed as

$$N(n,s,d) = \left[\frac{1}{W} \sum_{m=-\infty}^{+\infty} [x_n(m,s,d) - x_n(m,s,\phi)]^2 \right]^{1/2} \qquad 3.3.3\text{-}1$$

where $x_n(w,s,d)$ and $x_n(w,s,\phi)$ are the windowed distorted signal and undistorted signal, respectively, as defined by Equation 3.3.1-1, and W is the window length. Note that the limits on m are really finite because of the windowing process. In the same terms, the signal energy is defined as

$$SG(n,s) = \left[\frac{1}{W} \sum_{m=-\infty}^{\infty} (x_n(m,s,\phi))^2 \right]^{1/2} \qquad 3.3.3\text{-}2$$

and the traditional SNR can be defined as

$$\text{SNR} = O(d) = 10 \log_{10} \left[\frac{\sum_{s=1}^{4} \sum_{n=1}^{N} (SG(n,s))^2}{\sum_{s=1}^{4} \sum_{n=1}^{N} (N(n,s,d))^2} \right] \qquad 3.3.3\text{-}3$$

where O(d) indicates this is an objective measure and the definitions of terms is the same as in Section 3.2.

The second class of measures of interest were "short time" or "framed" noise measurements. In this measurement, a frame by frame signal-to-noise ratio is computed, and then a global average is computed as usual from Equation 3.2-1. In this measurement,

$$F = 20[\log_{10} G(n,s,d)]^{\delta} \qquad 3.3.3\text{-}4$$

where

$$\log_{10} [1 + G^2(n,s,d)] = \log_{10} \left[1 + \frac{SG^2(n,s,d)}{N^2(n,s,d)} \right] \qquad 3.3.3\text{-}5$$

and δ is a parameter for study. These short time signal-to-noise ratios have recently been shown to be more highly correlated with subjective results than traditional SNR measurements [3.14],[3.15].

3.4 Frequency Variant Objective Measures

One of the major hypotheses of the study was that, since it is well known that the perception of sound in humans is a frequency variant process, then frequency variant objective measures could be expected to perform better as predictors of subjective results than objective measures which are frequency invariant. One method of testing this hypothesis has already been discussed in the section on simple spectral distance measures. This was the technique of weighting the spectral distance measure by a function of the spectrum of the undistorted speech (see Equations 3.3.1.2-3 and 3.3.1.2-5). This section offers a different approach to frequency weighting, an approach in which the frequency weights are set so as to give maximum correlation between the objective measures and the subjective measures.

The analysis technique can be described as follows. First, a frequency sampled objective measure is defined. In this study, two such measures, spectral distance and short time banded signal-to-noise, were used. These two measures will be described in detail below. Let there be B frequency bands in the analyses. Then for each distortion, B different objective measures, $O_b(d)$, where b is the band index and d is the distortion index, are computed. In general, the subjective results for distortion d may be estimated by a linear sum of the banded objective measures by

$$\hat{S}(d) = \sum_{b=1}^{B} C(b)\ O_b(d) + C(0) \qquad 3.4\text{-}1$$

where S(d) is the estimate of the subjective measure S(d) and C(b) are a

set of unknown constants. The error between the true subjective result and the estimated subjective results is given by

$$E(d) = S(d) - \hat{S}(d). \qquad 3.4\text{-}2$$

Now, if the C(b), b = 0,...,B are chosen to minimize the squared error, then a maximum correlation between S(d) and $\hat{S}(d)$ is achieved. This minimization results in a set of equations

$$\underline{\Phi}\,\underline{C} = \underline{p} \qquad 3.4\text{-}3$$

where

$$\underline{C} = \begin{bmatrix} C(0) \\ C(1) \\ \cdot \\ \cdot \\ \cdot \\ C(B) \end{bmatrix} \qquad 3.4\text{-}4$$

$$\underline{p} = \begin{bmatrix} \sum_{d=1}^{D} S(d) \\ \sum_{d=1}^{D} S(d)\, O_1(d) \\ \cdot \\ \cdot \\ \cdot \\ \sum_{d=1}^{D} S(d)\, O_B(d) \end{bmatrix} \qquad 3.4\text{-}5$$

where $\phi(m,n)$, the (m,n) entry of the matrix $\underline{\phi}$, is given by

$$\phi(m,n) = \sum_{d=1}^{D} O_m(d)O_n(d) \qquad 3.4\text{-}6$$

where D is the total number of distortions considered. Clearly, an optimal set of values for C(b)'s may be obtained in this way for any set of distortions in the data base.

Several points should be discussed here. First, the correlation coefficients obtained between $\hat{S}(d)$ and S(d) after the C(b)'s have been found must be considered a limit on the correlation obtained by weighted frequency analysis. This, of course, is because the data itself is being used to compute both the correlations and the weights. Second, since for many of the distortions in the distorted data base the banded distortions are highly correlated with one another, the results of this analysis cannot be considered as a direct estimate of the underlying optimal physical weights. This is the reason that a large subset of the distorted data base is made of frequency banded distortions. Estimates based on this subset would have more universal validity than those taken across the entire data base. Finally, since the optimization of Equation 3.4-3 may be done against any of the different parametric subjective results (see Chapter 2), these measures may be "tuned" to predict specific parametric subjective results as well as isometric subjective results. Since many of the parametric subjective results are frequency variant in nature, such tuning should be very effective.

3.4.1 Banded Spectral Distance Measures

One of the two types of frequency variant distance measures considered is the frequency banded spectral distance measure. From Section 3.3, recall that in frequency invariant measures, the frequency index, $\theta_\ell = \frac{\pi \ell}{L}$ for $\ell = 0,\ldots,L$. This is clearly a one band analysis. For a B band analysis, the total frequency band (π radians) is divided into B sub-bands by

$$\theta_\ell = \frac{\pi \ell}{L} \qquad \frac{\theta_{b-1} L}{\pi} \leq \ell < \frac{\theta_b L}{\pi} \qquad 3.4.1\text{-}1$$

where θ_b is the upper band limit for $b^{\underline{th}}$ band. In this study, B was normally equal to 6.

To measure the banded spectral distance measure, the values for $O_b(d)$ were computed by the same techniques as discussed in Section 3.3.1 but using the reduced bands given by Equation 3.4.1-1. In this analysis, two types of spectral normalizations were computed. First, the spectra, $V(n,s,d,\theta)$, were normalized to have an area of one across the entire band, as before. Second, the spectra were normalized to have an area of one in each individual band. Since this second method gives a better fit to the overall spectrum, it was expected that it would give better correlation results.

3.4.2 Banded Noise Measures

The frequency banded noise measures are the second class of frequency variant measures considered in this study. Like all noise measures,

these were only applied to the subset of the distorted data base for which noise measures are meaningful.

The computation of the banded noise is illustrated in Fig. 3.4.2-1. As can be seen, the noise is computed in the usual way and then the results are filtered into (usually) 6 separate bands. If the banded time signal is given by $x_b(m,s,d)$ and the windowed banded time signal is given by

$$x_{n,b}(m,s,d) = x_b(m,s,d)W(m-nI) \qquad 3.4.2\text{-}1$$

where W(m) is the window function and I is the frame interval as before, then the banded noise energy for the $n^{\underline{th}}$ frame of the $s^{\underline{th}}$ speaker of the $d^{\underline{th}}$ distortion is given by

$$N_b(n,s,d) = \left[\frac{1}{W}\sum_{m=-\infty}^{+\infty}(x_{n,b}(m,s,d) - x_{n,b}(n,s,\phi))^2\right]^{1/2} \qquad 3.4.2\text{-}2$$

where, as before, the limit on m is really finite. The banded signal energy, $SG_b(n,s)$, is given by

$$SG_b(n,s) = \left[\frac{1}{W}\sum_{m=-\infty}^{+\infty}x_{n,b}^2(m,s,\phi)\right]^{1/2} \qquad 3.4.2\text{-}3$$

In this context then, the banded short time objective measure is computed

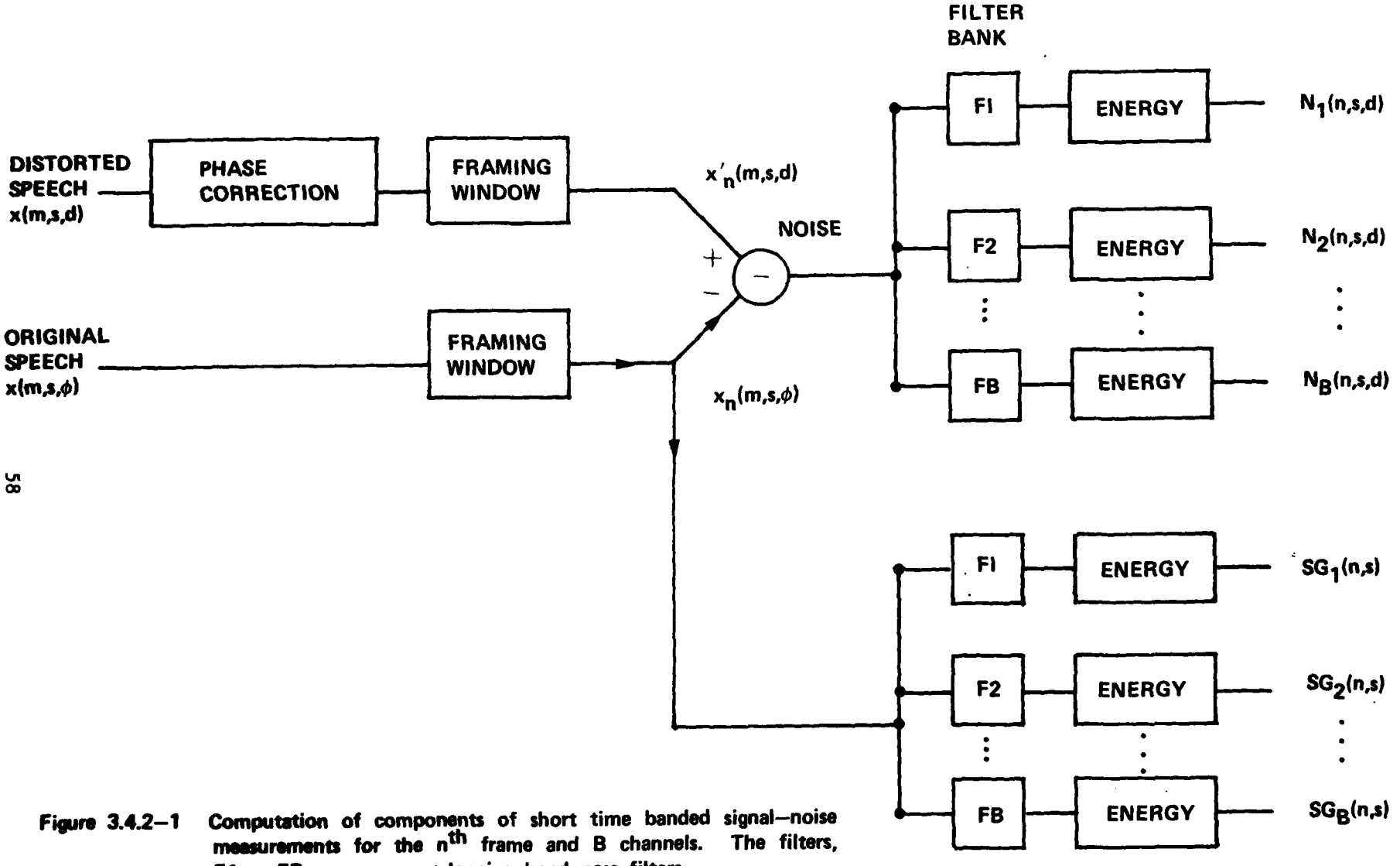


Figure 3.4.2–1 Computation of components of short time banded signal–noise measurements for the n^{th} frame and B channels. The filters, F1 – FB, are non–overlapping band pass filters.

from

$$F_b = 20[\log_{10} G_b(n,s,d))]^{\delta} \qquad 3.4.2\text{-}4$$

where

$$\log_{10}[1 + G_b^2(n,s,d)] = \log_{10}\left[1 + \frac{SG_b^2(n,s)}{N_b^2(n,s,d)}\right] \qquad 3.4.2\text{-}5$$

as before. In these studies, δ is a parameter for study.

3.5 The Composite Measures

The composite measures studied as part of this work were all taken to be linear combinations of groups of simple measures or frequency variant measures. The procedure in identifying and testing the composite measures was as follows. First, choose a set of candidate objective measures which have relatively high correlation with the subject results, and which are judged to be measuring different objective quantities. This measure will be designated $O_p(d)$, where this is the $p^{\underline{th}}$ measure of the distortion d. Second, rank these measures according to their estimated correlation with the subjective data base. Third, study all possible measures which are sums of two objective measures, i.e.

$$O(d) = g(1)O_{p_1}(d) + g(2)O_{p_2}(d) \qquad p_1 \neq p_2 \qquad 3.5\text{-}1$$

where g(1) and g(2) are unknown constants. Using least squares analysis (see section 3.4), choose the g(1) and g(2) for each combination which produces the highest correlation with the subjective data base. Fourth, study all measures which are combinations of 3,4,...,p objective measures using least squares (maximum correlation) analysis. Finally, within each group (1,2,...,p measures), rank the objective measures by their correlation coefficients.

This analysis produces the optimal, in a least squares sense, objective measure which can be constructed from the original p measures for a 1 term, 2 term, 3 term,..., and p term composite linear objective measures. This p term analysis can be thought of as a limit on the correlation obtainable from these measures. At each level, the measure with the highest correlation can be thought of as a limit on obtainable correlation for that number of terms. The level to level improvement supplies information as to the expected gain derivable from including additional measures as part of the composite measures. Fially, the weighting factors, g(k), form a vehicle for tuning these composite meassures to effectively predict parametric subjective results.

REFERENCES

3.1. T. P. Barnwell, A. M. Bush, R. M. Mersereau, and R. W. Schafer, "Speech Quality Measurement," Final Report, DCA Contract No. RADC-TR-78-122, June 1977.

3.2. B. S. Atal and M. R. Schroeder, "Optimizing Predictive Coders for Minimum Audible Noise," Proceedings of ICASSP, Washington, DC, April 1979.

3.3. T. P. Barnwell, "Objective Measures for Speech Quality Testing," JASA, December 1979.

3.4. J. L. Flanagan, Speech Analysis, Syntheses and Perception, Second Edition, Springer-Verlag, New York, 1972.

3.5. F. Itakura and S. Sarto, "On the Optimum Quantization of Feature Parameters in the PARCOR Speech Synthesizer," Proc. 1972 Con. Speech Comm. Process., 1972.

3.6. B. S. Atal and S. L. Hanover, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," JASA, 50, 1971.

3.7. J. D. Markel, "Digital Inverse Filtering--A New Tool for Format Trajectory Estimation," IEEE Trans. Audio ELECTROACOUSTICS, AU-20, 1972.

3.8. A. V. Oppenheim and R. W. Schafer, "Homomorphic Analysis of Speech," IEEE Trans. Audio ELECTROACOUSTICS, AU-16, 1968.

3.9. A. V. Oppenheim, "Speech Analysis-Synthesis System Based on Homomorphic Filtering," JASA, Vol. 45, 1969.

3.10. A. H. Gray, Jr. and J. D. Markel, "Distance Measures for Speech Processing," IEEE Trans. ASSP, Vol. 24, 1976.

3.11. F. Itakura, "Minimum Prediction Residual Principal Applied to Speech Recognition," IEEE Trans. ASSP, Vol. 23, 1975.

3.12. M. R. Sambur and N. S. Jayant, "LPC Synthesis Starting from White Noise Corrupted or Differentially Quantized Speech," Proc. of ICASSP, Philadelphia, PA, April 1976.

3.13 B. J. McDermott, C. Scagliola, and D. J. Goodman, "Perceptual and Objective Evaluation of Speech Processed by Adaptive Differential PCM," Proc. of ICASSP, Tulsa, OK, April 1978.

3.14. P. W. Noll, "Adaptive Quantization in Speech Coding Systems," IEEE Int. Zurich Sem. on Dig. Comm., October 1976.

3.15. P. Mermelstein, "Evaluation of Two ADPCM Coders for Toll Quality Speech Transmission," JASA, to be published, December 1979.

CHAPTER 4

THE DISTORTED DATA BASE

This chapter describes in detail the contents of the "distorted data base." As was discussed in the introduction, the undistorted data base consisted of four sets of twelve sentences, each set spoken by a different speaker, and each of which was band limited to 3.2 kHz and sampled to 12 bits resolution at 8 kHz. The total duration of the twelve sentence sets were adjusted to be 49.152 sec., or 393216 time samples, for each set. There were three male speakers, CH, LL, and RH, and one female speaker, JS.

A total of 264 "distortions" were identified and applied to the undistorted data base (see Table 4-1). The distortions can be roughly divided into two types: "coding" distortions, which are simulations of digital coding systems; and "controlled" distortions, in which some specific perceptually relevant distortion is applied to the speech. All distortions were applied digitally using the Georgia Tech Minicomputer Based Digital Signal Processing Laboratory [4.1]. The 264 distortions are subdivided into 44 types of distortions and, within each type, there are six levels of distortion. The total length of the distorted sentences after preparation for subjective testing was over 17 hours, excluding anchors and probes.

Subjective testing was applied to the distorted data base using eleven four speaker DAM's (see Chapter 2). Each DAM tested four types of distortions for each of their six levels, giving 24 distortions per DAM. The contents of the individual runs is given in Table 4-2.

DISTORTIONS	NO. OF DISTORTIONS
Coding Distortion	
Adaptive PCM (APCM)	6
Adaptive Differential PCM (ADPCM)	6
CVSD	6
Adaptive Delta Modulator (ADM)	6
Adaptive Predictive Coding (APC)	6
Linear Predictive Coding (LPC)	6
Voice Excited Vocoder (VEV)	12
Adaptive Transform Coder (ATC)	6
	54
Controlled Distortions	
Additive Noise	6
Low Pass Filter	6
High Pass Filter	6
Band Pass Filter	6
Interruption	12
Clipping	6
Center Clipping	6
Quantization	6
Echo	6
	60
Frequency Variant Controlled Distortions	
Additive Colored Noise	36
Banded Pole Distortion	78
Banded Frequency Distortion	36
	150
TOTAL	264

Table 4-1. TOTAL SET OF DISTORTIONS IN THE DISTORTED DATA BASE.

CONTENTS OF THE INDIVIDUAL DAM RUNS

Run Number	Distortion	
1	Additive noise	(6)
	Low pass filter	(6)
	High pass filter	(6)
	Band pass filter	(6)
2	Interrupted	(12)
	Clipping	(6)
	Center clipping	(6)
3	Colored noise	(24)
4	Colored noise	(12)
	APCM	(6)
	ADPCM	(6)
5	Banded freq. dist.	(24)
6	Banded freq. dist.	(6)
	Banded pole dist.	(18)
7	Banded freq. dist.	(6)
	Banded pole dist.	(18)
8	Banded pole dist.	(24)
9	Banded pole dist.	(18)
	Echo	(6)
10	ADM	(6)
	CVSD	(6)
	APC	(6)
	Quantization	(6)
11	LPC	(6)
	VEV	(12)
	ATC	(6)

Table 4-2. Contents of the Individual DAM Runs.

The remainder of this chapter will be devoted to describing the individual distortions.

4.1 The Coding Distortions

In all, there were nine types of coding distortions used in this study, resulting in a total set of 60 distortions. In all cases, the coding distortions were simulated and were designed to be zero phase if possible. They were always at least designed so that the distorted speech would have frame by frame synchronization with the undistorted speech.

4.1.1 Simple Waveform Coders

In this study, there were four systems which were classed as "simple" waveform coders: Continuously Variable Slope Delta Modulator (CVSD): Jayant's [4.1] Adaptive Delta Modulator (ADM); Adaptive Pulse Code Modulation (APCM); and Adaptive Differential Pulse Code Modulation (ADPCM). All of these systems can be thought of as special cases of the general adaptive waveform coding system illustrated in Fig. 4.1.1-1. In all cases, the interpolater, where used, was implemented using zero phase FIR interpolation filters implemented with FFT techniques, as was the decimation. The "channel simulation" shown in these systems was always only capable of introducing random bit errors at fixed rates and simulated no other characteristic of a real channel.

4.1.1.1 CVSD

The CVSD is a delta modulator, so that the quantizer is always a two level quantizer and the coder is a one bit coder. The main feature of the CVSD is in the way it computes $\delta(n)$ (see Figure 4.1.1-1). Since $\xi(n)$ is the output of a one bit quantizer, it may be thought of as a series of ± 1's.

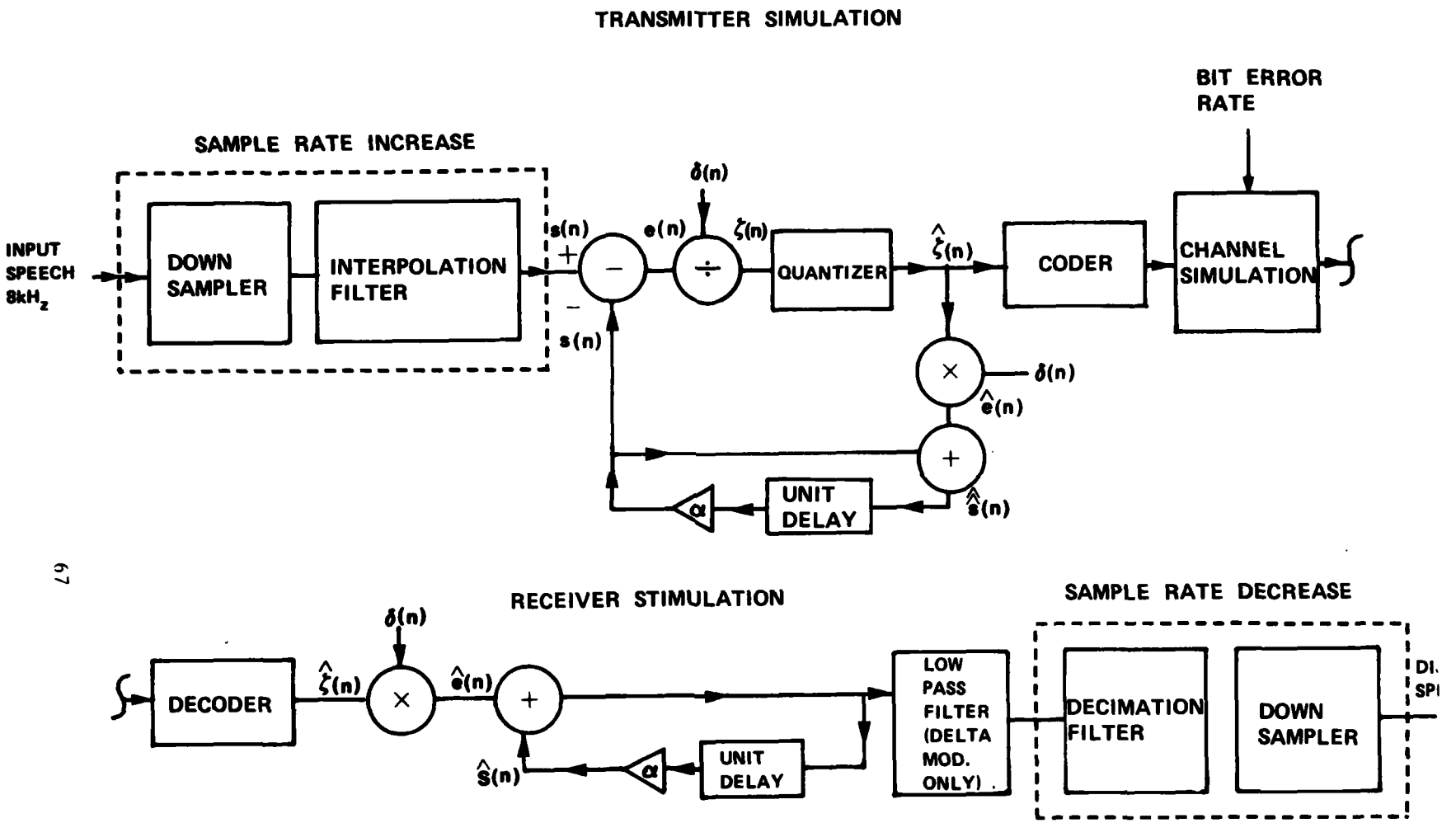


Figure 4.1.1–1 General System for Describing Waveform Coders

$\delta(n)$ is computed as

$$\delta(n) = \beta\delta(n-1) + \Delta(n) \qquad 4.1.1.1\text{-}1$$

where $\Delta(n)$ is equal to one of two constants depending on whether all of the last three values of $\xi(n)$ were equal to one another or not. So

$$\Delta(n) = \begin{cases} A & \text{if last 3 } \xi(n) \text{ equal} \\ B & \text{if last 3 } \hat{\xi}(n) \text{ not equal} \end{cases} \qquad 4.1.1.1\text{-}2$$

B is known as the "minimum step size" for CVSD. The corresponding maximum step size is given by $(\frac{1}{1-\beta})A$.

CVSD is hence characterized by five features: the input speech sampling rate; the value of the predictor parameter, α; the value of the "step integrator" parameter, β; the value of the minimum step size, B; and the value of A. A is usually not given, but is rather represented as an "expansion ratio," which is the maximum step size divided by the minimum step size, giving $(\frac{1}{1-\beta})\frac{A}{B}$.

In terms of its basic parameters, the CVSD systems used in this study are summarized in Table 4.1.1.1-1.

4.1.1.2 ADM

The adaptive delta modulator used in this study was essentially suggested by Jayant [4.2]. Like CVSD, the ADM is a delta modulator, so the different data rates are controlled by the interpolation process, the quantizer is a one bit quantizer, and the coder is a one bit coder. For this delta modulator,

$$\delta(n) = \Delta(n)\delta(n-1) \qquad 4.1.1.2\text{-}1$$

	Predictor Constant (α)	Step Size Integration (β)	Minimum Step Size (B)	Expansion Ratio	Bit Rate
1	.86	.9922	10	166	8 KBPS
2	.9696	.9922	10	166	12 KBPS
3	.98	.9922	10	166	16 KBPS
4	.99	.9922	10	166	24 KBPS
5	.995	.9922	10	166	32 KBPS
6					original

Table 4.1.1.1-1. Parameters for CVSD.

where $\Delta(n)$ takes on one of two values: "A" where $\hat{\xi}(n)$ and $\hat{\xi}(n-1)$ are equal; and "B" when they are not. In general, A is greater than one and B is less than one. For this study,

$$A = 1/B \qquad 4.1.1.2\text{-}2$$

The ADM is hence characterized by only three parameters: the input speech sampling rate; the value of the predictor parameters, α; and the value of the quantizer control parameter, A. In terms of these parameters, the ADM distortions used in this study are summarized in Table 4.1.1.2-1.

4.1.1.3 APCM

APCM has three main characteristics: first, it uses a multilevel quantizer; second, it operates at the Nyquist rate, and hence the interpolation and decimation filters are not used; and third, it has no prediction loop, i.e., $\alpha = 0$. The quantizer control sequence, for this study, was controlled exponentially from

$$Z(n) = \beta\delta(n-1) + (1-\beta)|\hat{\xi}(n)| \qquad 4.1.1.3\text{-}1$$

This can be thought of as an exponentially integrated estimation of the energy in the quantized error signal, $\hat{\xi}(n)$. From this,

$$\delta(n) = \frac{Q4}{N} Z(n) \qquad 4.1.1.3\text{-}2$$

where Q is a control parameter and N is the number of levels in the quantizer. This realization is, therefore, completely controlled by three parameters: the quantizer integration factor, β; the quantizer

	Predictor Constant (α)	Single Bit Multiplier (A)	Bit Rate
1	.86	1.1	8 KBPS
2	.90	1.06	12 KBPS
3	.96	1.03	16 KBPS
4	.98	1.03	24 KBPS
5	.99	1.03	32 KBPS
6			original

Table 4.1.1.2-1. Parameters for Adaptive Delta Modulator (ADM)

multiplier, Q; and the number of levels, N. In terms of these parameters, the APCM distortions used in this study are given in Table 4.1.1.3-1.

4.1.1.4 ADPCM

The ADPCM used in this study was exactly the same as the APCM previously described except the value of α was not zero. The operation of this system is hence characterized by four parameters: the quantizer integration factor, β; the quantizer multiplier, Q; the number of quantizer levels, N; and the feedback parameter, α. In terms of these parameters, Table 4.1.1.4-1 describes the ADPCM distortions used in this study.

4.1.2 The LPC Vocoder

The operation of the LPC vocoder used in this study is illustrated in Figure 4.1.2-1. This procedure is a framed analysis and is characterized by a frame interval, I. At each frame interval the input speech, x(m,s,d), is windowed, as before, to give

$$x_n(m,s,d) = x(m,s,d)W(m-nI) \qquad 4.1.2\text{-}1$$

where W(m) is a window function of length W, n is the frame number, m is the time index, s is the speaker, and d is the distortion. For this study, a Hamming window was used. From this, a set of autocorrelation functions is estimated from

$$R(k) = \sum_{m=-\infty}^{+\infty} x_n(m,s,d)x_n(m-k,s,d) \qquad 4.1.2\text{-}2$$

	Quantizer Integration (β)	Quantizer Multiplier (Q)	# of Levels (N)	Bit Rate (BPS)
1	.92	1	3	12679
2	.92	1	5	18575
3	.92	1	7	22453
4	.92	1	9	25359
5	.92	1	11	27675
6	.92	1	13	29603

Table 4.1.1.3-1. Parameter for Adaptive Pulse Code Modulation (APCM).

	Predictor Constant (α)	Quantizer Integration (β)	Quantizer Multiplier (Q)	# of Levels	Bit Rate (BPS)
1	.9	.92	1	3	12679
2	.9	.92	1	5	18575
3	.9	.92	1	7	22458
4	.9	.92	1	9	25359
5	.9	.92	1	11	27675
6	.9	.92	1	13	29603

Table 4.1.1.4-1. Parameter for Adaptive Differential Pulse Code Modulation (ADPCM).

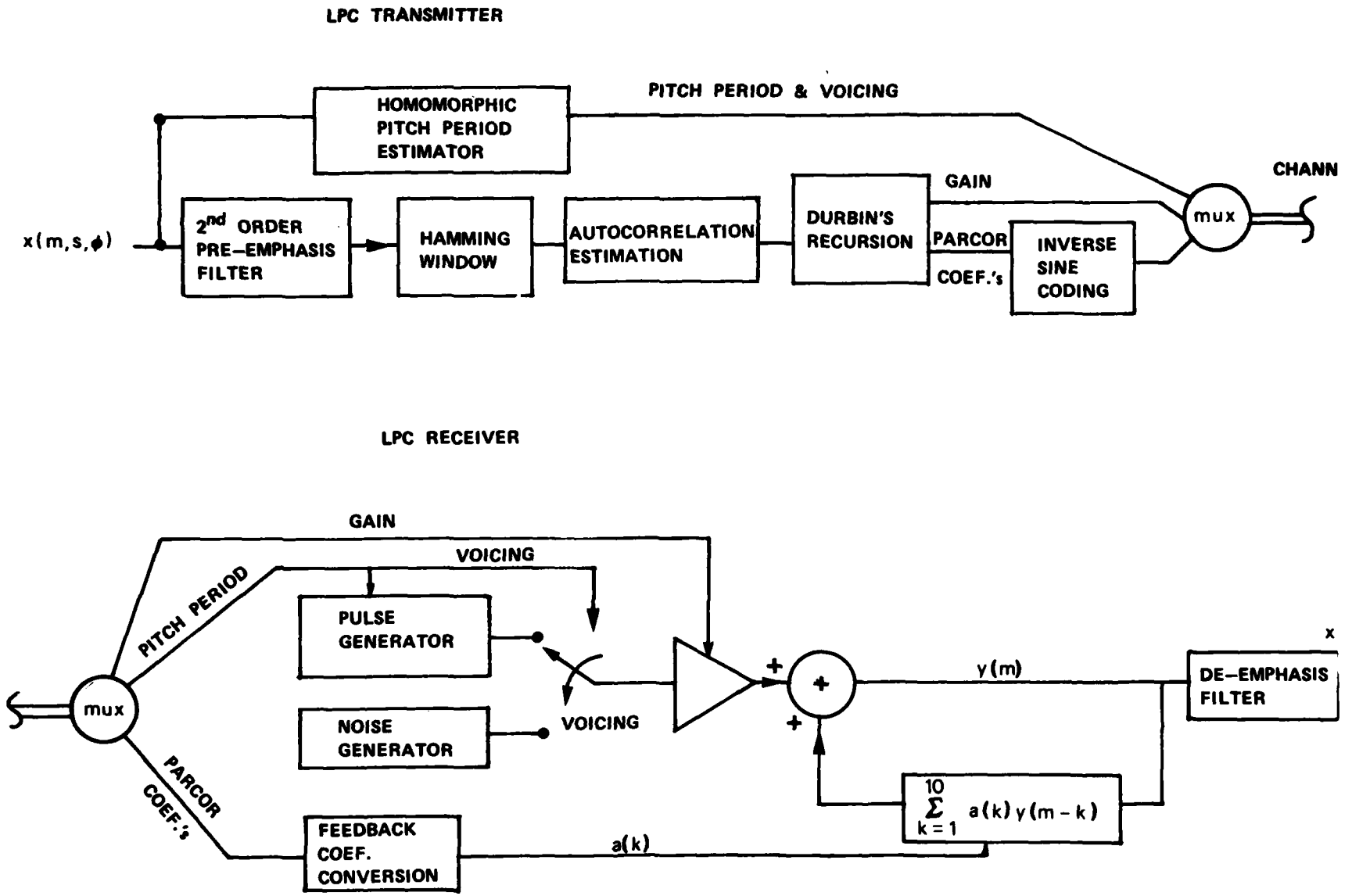


FIGURE 4.1.2–1 LINEAR PREDICTIVE CODER (LPC) SIMULATED TO FORM THE LPC CODING DISTORTION

Using the well known Durbin's recursion (see 3.3.1.1) a set of feedback coefficients, $\alpha(1)...a(10)$, a set of PARCOR coefficients, $K(1)...K(10)$, and a gain, given by

$$G = [R(0) - \sum_{k=1}^{10} a(k)R(k)]^{1/2} \qquad 4.1.2\text{-}3$$

is computed.

The set of PARCOR coefficients are an equivalent set of parameters to the feedback coefficients which may be interchanged by the recursions

$$a^1(1) = -K(1)$$

$$a^n(k) = a^{n-}(k) + K(n)a^{n-1}(n-k) \qquad 4.1.2\text{-}4$$

$$a^n(k) = -K(n) \qquad k = 1,2,...,n-1$$

and

$$b^N(k) = -a(k)$$

$$K(N) = b^N(N)$$

$$K(n) = b^n(n)$$

$$b^{n-1}(k) = (b^{n-1}(k)-K(n)b^{n-1}(n-k))/1-K^2(n) \quad k=1,...n-1 \qquad 4.1.2\text{-}5$$

Since the spectral sensitivity to quantization errors increases when the

PARCOR coefficients have values close to ± 1, the inverse sine transform of the parameters is used [4.3].

The pitch detector used is a form of "Homomorphic" or Cepstrum" pitch detector [4.4], [4.5]. The pitch and voicing output from the pitch detector is multiplexed in with the vocal tract information for transmission.

There are four parameters which characterize the LPC vocoder distortion. They are the window length, W; the number of bits per frame for the PARCOR coefficients; the frame interval, I; and the pitch and gain bits. The LPC distortions used in this study are described in terms of these parameters in Table 4.1.2-1.

4.1.3 The Adaptive Predictive Coder (APC)

The operation of the APC used in this study is illustrated in Figure 4.1.3-1. In this system, the first step is that a framed LPC analyzer is applied to the input speech waveform. The LPC analyzer is the same as that described in section 4.1.2, and produces a vector of feedback coefficient, a(k) for k = 1,...,10. This information is coded to some fixed bit rate using "inverse sine" PARCOR quantization [4.3] and then used to control a time varying prediction filter with the Z transform

$$P(Z) = 1 - \sum_{k=1}^{10} a(k)Z^{-k} \qquad 4.1.3\text{-}1$$

The {a(k)} coefficients are also transmitted to the receiver. The adaptive predictor, inside the prediction loop, is then used to estimate the input sequence x(m). The error signal, e(n), between the input sequences and the output of the predictor is then quantized by an adaptive quantizer

	Window Length (msec)	Bits/ Frame (vocal tract)	Pitch & Gain	Frame Interval (msec)	Bit Rate (BPS)
1	30	unquantized	7	15	--
2	30	58	7	15	4333
3	30	48	7	15	3666
4	30	38	7	15	3000
5	30	29	7	15	2400
6	30	20	7	15	1800

Table 4.1.2-1. Parameters for the LPC Vocoder.

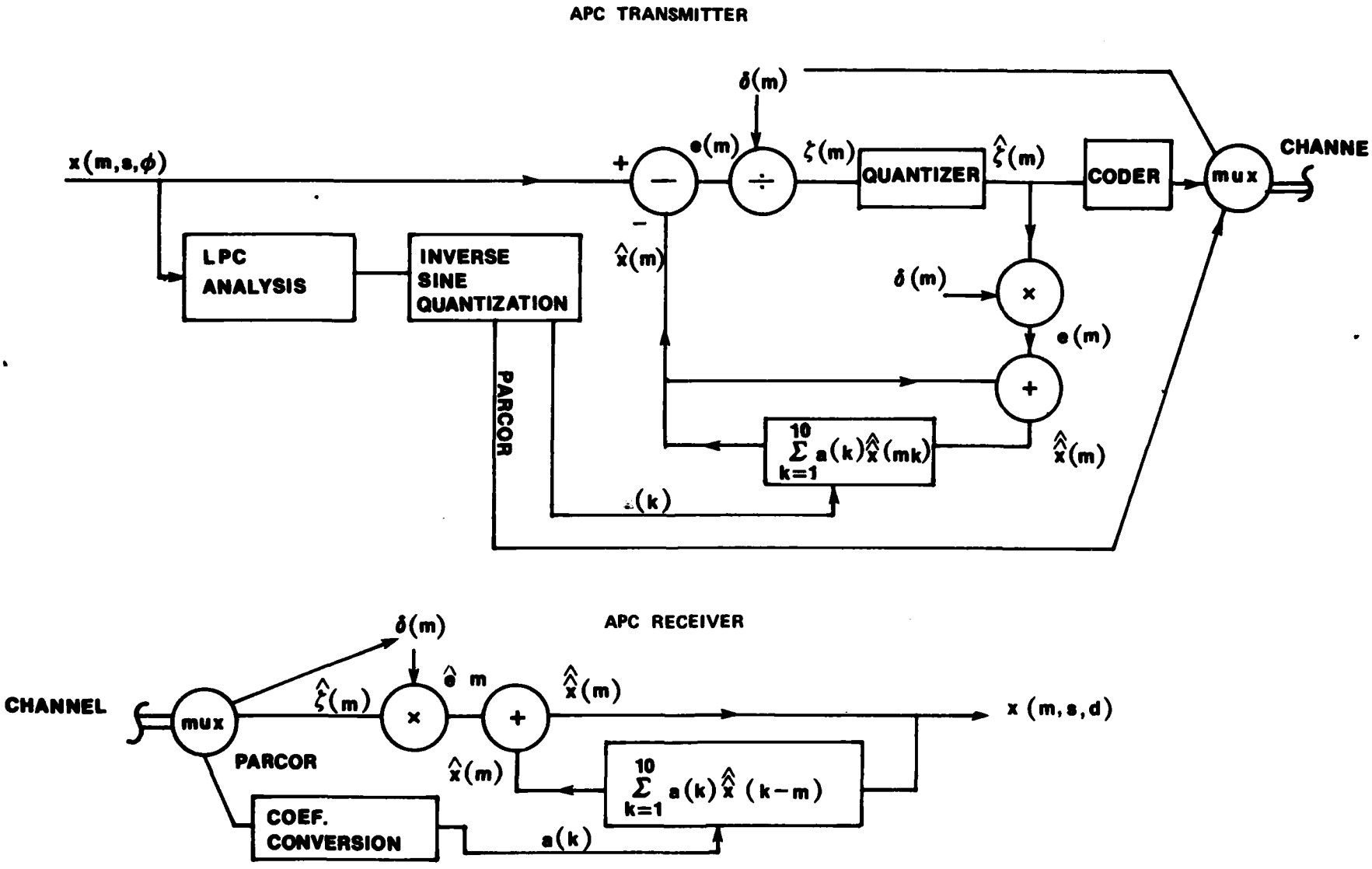


FIGURE 4.1.3–1 THE ADAPTIVE PREDICTIVE CODING SYSTEM USED AS PART OF THE CODING DISTORTION STUDY

consisting of a AGC followed by a fixed quantizer. In this simulation $\delta(n)$ was taken to be the "look ahead" frame energy average, given by

$$\delta(n) = \left[\frac{1}{W} \sum_{m=\infty}^{+W} X_n^2 (m,s,d) \right]^{1/2} \frac{Q4}{N} \qquad 4.1.3\text{-}2$$

where W is the window length, Q is a quantizer control parameter, and N is the number of levels in the uniform quantizer.

The total operation of this APC is then characterized by five factors: the number of levels in the quantizer N; the frame rate; the window length; the number of bits per frame in the predictor coding; and the quantizer control factor, Q. In terms of these parameters, the APC distortion used for this study is given in Table 4.1.3-1.

4.1.4 The Voice Excited Vocoder (VEV)

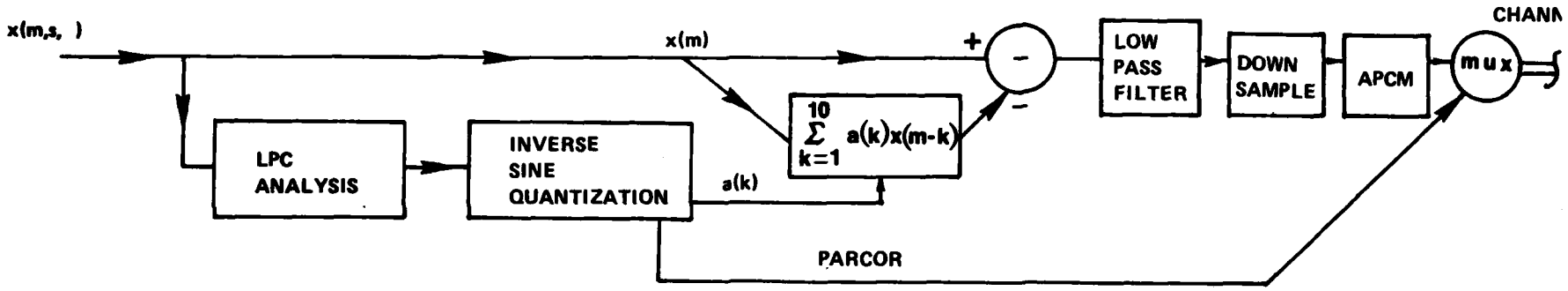
The voice excited vocoder used in this study is illustrated in Figure 4.1.4-1. Its operation is essentially similar to the APC described in section 4.1.3 except for the following features. Instead of sending the entire residual signal, $\hat{\xi}(n)$, a low passed version of this signal is sent. There is some data rate compression gained by coding and down sampling this low passed signal to the Nyquist rate appropriate to its bandwidth. At the receiver, the excitation function is recreated by using the base band, where appropriate, and using a full wave rectification and LPC flattening to regenerate the higher frequency.

The VEV vocoder simulated here is characterized by five parameters: the frame interval, I; the window length, W; the ADPCM transmission rate; the voice band bandwidth; and the vocal tract parameter bit rate. Table 4.1.4-1 described the VEV distortions used in this study as a function of these parameters.

	Window Length	Bits/ Frame	Level (N)	Frame Interval	Bit Rate (BPS)
1	30	unquantized	3	15	--
2	30	58	3	15	15867
3	30	48	3	15	15200
4	30	38	3	15	14533
5	30	29	3	15	13933
6	30	20	3	15	13333

Table 4.1.3-1. Parameters for the Adaptive Predictive Coder (APC).

VOICE EXCITED VOCODER TRANSMITTER



VOICE EXCITED VOCODER RECEIVER

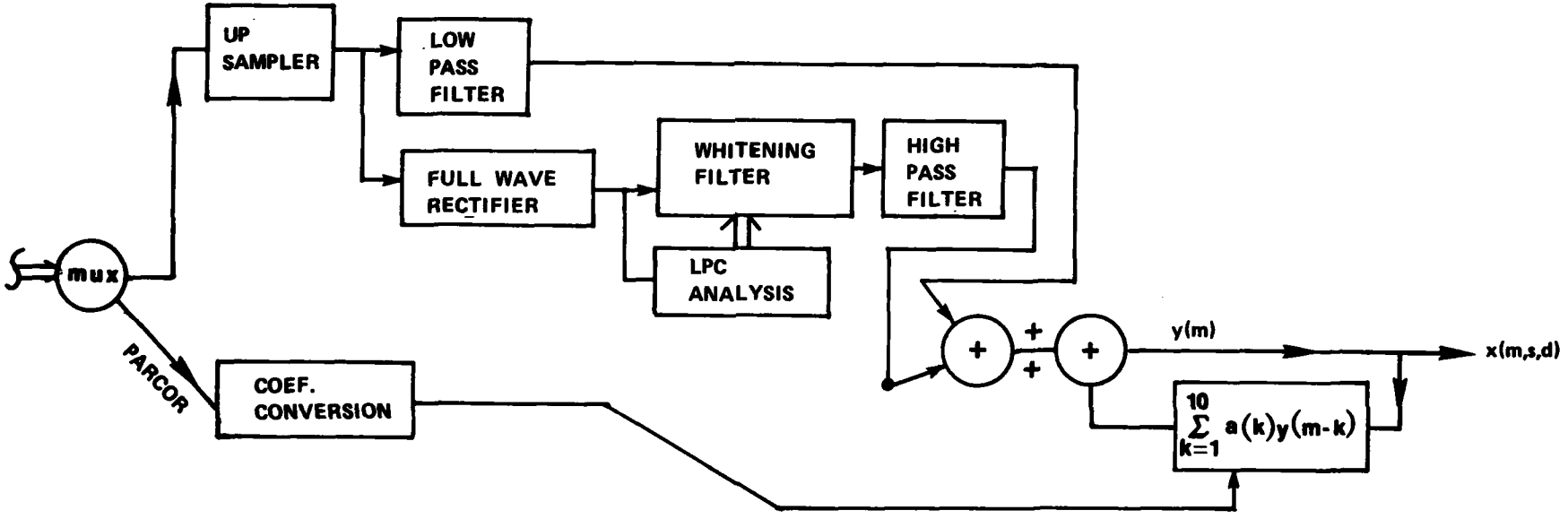


FIGURE 4.1.4–1 VOICED EXCITED VOCODER

	FRAME INTERVAL (I) (msec)	WINDOW LENGTH (W) (msec)	ADPCM RATE (BPS)	VOICE BAND (HZ)	VOCAL TRACT DATA	TOTAL DATA RATE
1	15	30	5615	1000	3867	9482
2	15	30	5615	1000	3200	8815
3	15	30	5615	1000	2533	8148
4	15	30	5615	1000	1933	7548
5	15	30	5615	1000	1333	6948
6	15	30	5615	1000	1000	6615
7	15	30	7400	1000	3867	11267
8	15	30	7400	1000	3200	10600
9	15	30	7400	1000	2533	9933
10	15	30	7400	1000	1933	9333
11	15	30	7400	1000	1333	8733
12	15	30	7400	1000	1000	8400

Table 4.1.4-1. Parameters for the Voice Excited Vocoder (VEV).

4.1.5 Adaptive Transform Coding (ATC)

Adaptive transform coding is a relatively new coding technique as applied to speed [4.6], [4.7], and one that has been shown to have great promise. In this study, it was not desired to produce high quality ATC speech, because that was still a subject of research at the time these distortions were chosen. Rather it was to include in the data base a distortion which was qualitatively "like" that produced by ATC.

The ATC coding system used in this study is illustrated in Figure 4.1.5-1. First, the speech is windowed to 256 samples using a rectangular window and a frame interval of 256 points also. Each windowed speech sample is then both transformed using the DCT and analyzed using LPC analysis. An approximate spectrum is computed from the LPC analyzer from

$$V(\theta_\ell) = \left| \frac{1}{1 - \sum_{k=1}^{10} a(k) e^{-jk\theta_\ell}} \right| \qquad 4.1.5\text{-}1$$

and then the levels are allocated at spectral sample θ_ℓ, $0<\ell<255$, by

$$\text{levels}\ (\theta_\ell) = (\text{TOTAL LEVELS}) \cdot V(\theta_\ell) \qquad 4.1.5\text{-}2$$

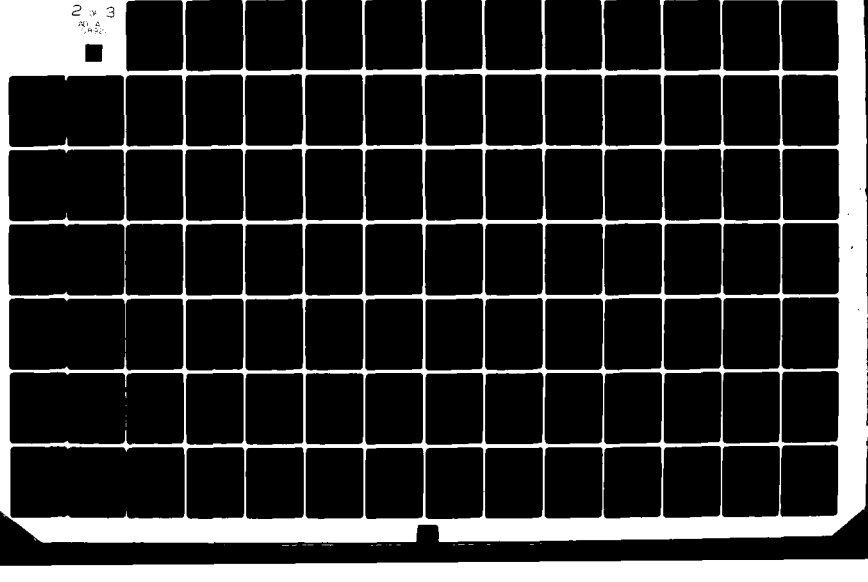
(recall that $\sum_{\ell=o}^{255} V(\theta_\ell) = 1$), where if B is the total bits allocated, then

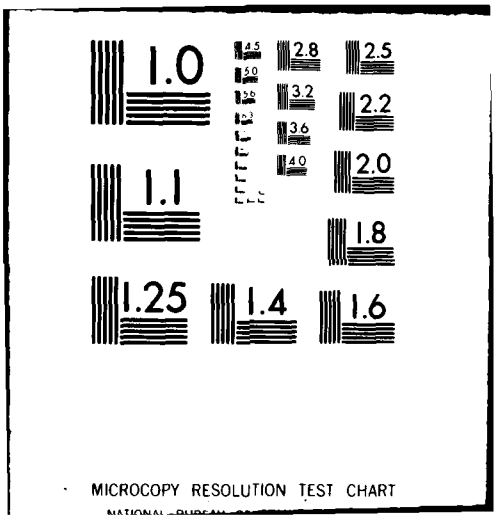
$$\text{TOTAL LEVELS} = 2^B \qquad 4.1.5\text{-}3$$

The individual quantizers are uniform with a range, $r(\ell)$ given by

$$-GV(\theta_\ell) < r(\ell) < GV(\theta_\ell) \qquad 4.1.5\text{-}2$$

AD-A089 210 GEORGIA INST OF TECH ATLANTA SCHOOL OF ELECTRICAL EN--ETC F/G 17/2
AN ANALYSIS OF OBJECTIVE MEASURES FOR USER ACCEPTANCE OF VOICE --ETC(U)
SEP 79 T P BARNWELL, W D VOIERS DCA100-78-C-0083
UNCLASSIFIED E21-659-78-T8-1 NL




1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
MICROCOPY RESOLUTION TEST CHART

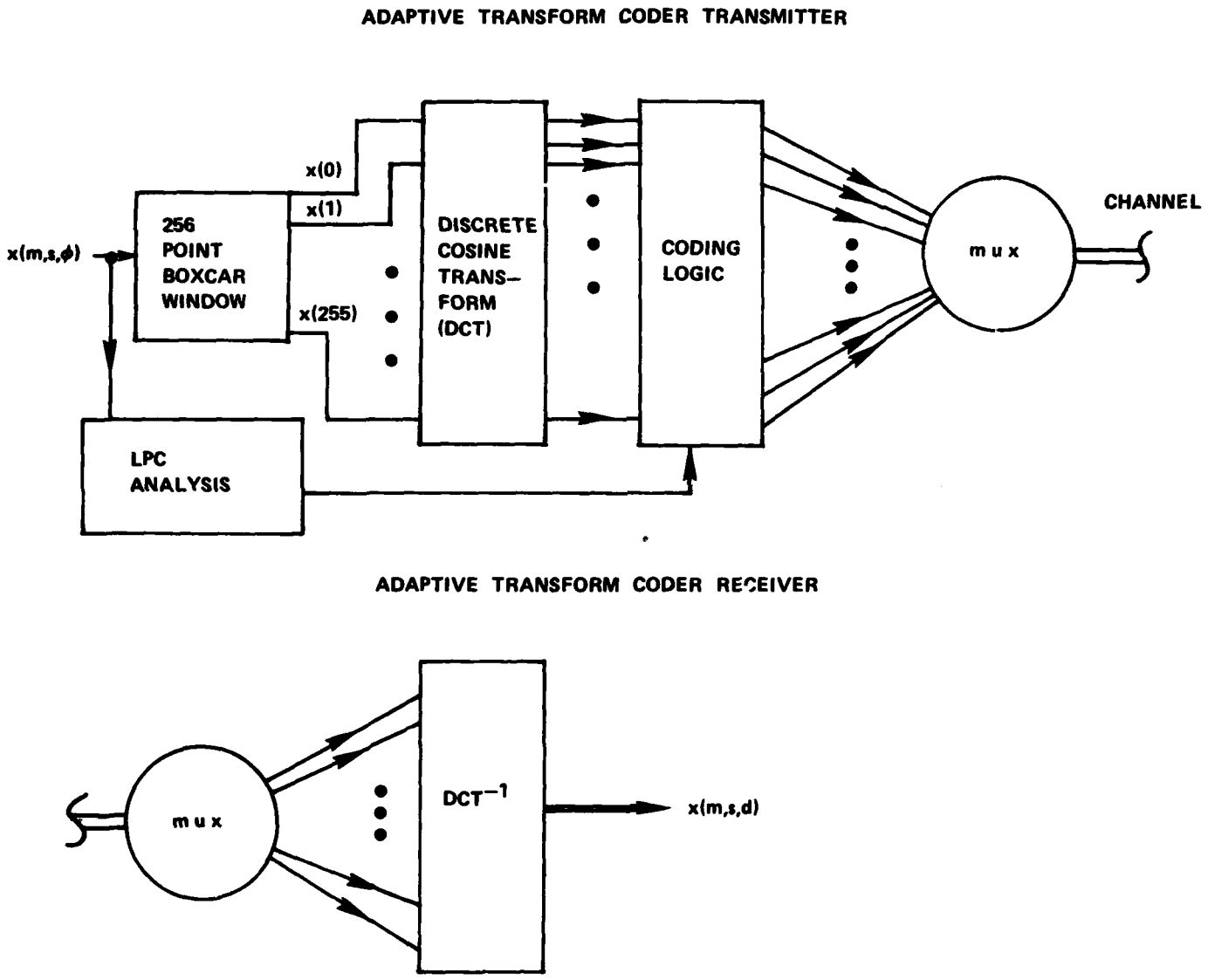


Figure 4.1.5–1 ADAPTIVE TRANSFORM CODER USED FOR THE DISTORTED DATA BASE

where G, the gain, is given by equation 4.1.2-3.

The operation of this transform coder is characterized by 4 parameters: The frame interval and window length, which must be the same; the order of the LPC; the LPC vocal tract parameter bits per frame; and the transform coder bits per frame, B. The distortions used in this ATC system are summarized in terms of these parameters in Table 4.1.5-1.

4.2 The Controlled Distortions

A large portion of distortions used in this study were not explicit coding distortions, but were "controlled" distortions. These distortions were included for one of two reasons. Either they were considered to be examples of specific types of subjectively relevant distortions, or they were considerd to be one type of which occurs in coding distortion, but which does not occur in isolation.

A large portion of the controlled distortions are frequency variant distortions. These distortions are included for two reasons: first, they offer a measure of the subjective importance of different types of distortions when applied in different bands; and, second, they offer an environment in which the frequency variant objective measures will be relatively uncorrelated from band to band.

4.2.1 Simple Controlled Distortions

In this section, each of the non-frequency variant controlled distortions will be discussed separately.

4.2.1.1 Additive Noise

In the additive noise distortions, white Gaussian noise was added to each sample of the undistorted signal, i.e.,

	Window Length Frame Interval	LPC Order	LPC Bits/ Frame	Trans Bits/ Frame	Bit Rate (BPS)
1	256	10	4,333	15,667	20,000
2	256	10	3,666	12,334	16,000
3	256	10	3,000	9,000	12,000
4	256	10	2,400	8,600	11,000
5	256	10	1,800	7,800	9,600
6	256	10	1,500	6,500	8,000

Table 4.1.5-1. PARAMETERS FOR THE ADAPTIVE TRANSFORM CODER

$$x(m,s,d) = x(m,s,\phi) + A \cdot n(m) \qquad 4.2.1.1\text{-}1$$

where n(m) is a zero mean unit variance white noise sequence, and A is a multipicative constant. This distortion is well characterized by its signal-to-noise ratio (SNR) as shown in Table 4.2.1.1-1.

4.2.1.2 Filtering Distortions

There were three filtering distortions included: low pass filtering; high pass filtering; and band pass filtering. The filters were implemented digitally using recursive eliptical filters, i.e.,

$$x(m,s,d) = \sum_{k=0}^{K} b(k)x(m-k,s,\phi) + \sum_{k=1}^{K} a(k)x(m-k,s,d) \qquad 4.2.1.2\text{-}1$$

where K is the order of the eliptical filters. Table 4.2.1.2-1 gives the orders of the filters used along with the band limits for each distortion.

4.2.1.3 Interruptions

The interruption distortion was characterized by two numbers: a "keep" number, KP, and a "discard" number, DR. The interrupt distortion operated on frames of length KP + DR. Within in frame, the first KP samples were undisturbed, while the last DP were set to zero. Table 4.2.1.3 summarizes the interrupt distortions in this study.

4.2.1.4 Clipping

The clipping distortion is a nonlinear distortion given by

$$x(m,s,d) = \begin{cases} CL & |x(m,s,\phi)| \geq CL \\ x(m,s,\phi) & |x(m,s,\phi)| < CL \end{cases} \qquad 4.2.1.4\text{-}1$$

	Signal-to-Noise Ratio(DB)
1	30
2	24
3	18
4	12
5	6
6	0

Table 4.2.1.1-1. THE ADDITIVE NOISE DISTORTION

Low Pass Filters

	Order	Band Limit(HZ)
1	6	400
2	6	800
3	7	1,300
4	7	1,900
5	7	2,600
6	5	3,400

High Pass Filters

	Order	Band Limit
1	4	0
2	6	400
3	7	800
4	7	1,300
5	7	1,900
6	7	2,600

Band Pass Filter

	Order	Lower Band Limit	Upper Band Limit
1	6	0	400
2	9	400	800
3	9	300	1,300
4	9	1,300	1,900
5	9	1,900	2,600
6	11	2,600	3,400

Table 4.2.1.2-1. FILTER CHARACTERISTICS FOR RECURSIVE FILTERS USED FOR FILTER DISTORTION

	KP Keep Constant	DR Discard Constant
1	300	10
2	300	25
3	300	50
4	300	75
5	300	110
6	300	150
7	1,024	16
8	1,024	32
9	1,024	64
10	1,024	128
11	1,024	256
12	1,024	512

Table 4.2.1.3-1 "KEEP" AND "DROP" CONSTANTS FOR INTERRUPT DISTORTION

where the constant CL is called the clipping constant. The constant must be compared to the "maximum average energy," MAE, for an utterance, given by

$$MAE = MAX[E(m)\ ] \qquad 4.2.1.4\text{-}2$$

where E(m) is given by

$$E(m) = (1-\alpha)E(m-1)+ \alpha\, x(m,s,\phi) \qquad 4.2.1.4\text{-}3$$

where α is an exponential integration constant set to have a window length ~ 30 msec. For all the input sentences, the MAE was set to be .122 on a scale $-1 \leq x(m,s,d) \leq 1$. In these terms, the clipping constants for the clipping distortions are shown in Table 4.2.1.4-1.

4.2.1.5 Center Clipping

The center clipping distortion is a non-linear distortion given by

$$x(m,p,d) = \begin{cases} x(m,x,\phi) & |x(m,s,\phi)| \geq CN \\ 0 & |x(m,s,\phi)| < CN \end{cases} \qquad 4.2.1.5\text{-}1$$

where CN is the "center clipping constant." Table 4.2.1.5-1 gives the parameters for the distortion on the same scale as for clipping.

	Clipping Constant
1	.152
2	.076
3	.038
4	.0305
5	.0153
6	.0076

Table 4.2.1.4-1 CLIPPING CONSTANTS FOR CLIPPING DISTORTION

	Center Clipping Constant
1	.0019
2	.0038
3	.0076
4	.019
5	.038
6	.076

Table 4.2.1.5-1 CENTER CLIPPING CONSTANT FOR CENTER CLIPPING DISTORTION

4.2.1.6 Quantization Distortion

The quantization distortion is just a PCM system which is non-adaptive and which uses relatively coarse quantization. The quantizers used were always chosen to be linear and to cover a range of twice the maximum energy (see 4.2.1.4). The quantization distortion is described in terms of the number of levels in the quantizer and the associated bit rate in Table 4.2.1.6-1.

4.2.1.7 Echo Distortion

The Echo distortion was implemented by

$$x(m,s,d) = \frac{1}{2}\,[x(m,s,\phi)+x(m-EC,s,\phi)] \qquad 4.2.1.7$$

This is clearly not the only way to implement an echo, but the result is very clearly a subjective echo. The distortion is entirely characterized by the "echo delay," EC, and is described in Table 4.2.1.7-1.

4.2.2 Frequency Variant Controlled Distortions

This study included a total of three types of frequency variant controlled distortion. The first, the "additive colored noise," was designed to approximate waveform coder distortions in a frequency variant way. The second, called "pole distortion," was to approximate vocal tract modeling distortions in vocoders and APC's in a frequency variant way. Finally, the "banded waveform distortion" was designed to approximate the distortions found in ATC and adaptive subband coders in a frequency variant way.

	Number of Levels in Quantizer	Bit Rate
1	64	48,000
2	48	44,679.7
3	32	40,000
4	24	36,679.7
5	16	32,000
6	12	28,679.7

Table 4.2.1.6-1. QUANTIZATION DISTORTION PARAMETERS

	Echo Constant
1	10
2	50
3	100
4	200
5	500
6	1,000

Table 4.2.1.7-1. ECHO CONSTANT FOR THE ECHO DISTORTION

4.2.2.1 Additive Colored Noise

The additive colored noise system is illustrated in Figure 4.2.2.1-1. White Gaussian noise is first bandpass filtered into six bands giving an output signal $N_b(m)$, where b is the band number and m is the time index. Then the banded noise is added to the input speech using a noise constant, NC, giving

$$x(m,s,d) = x(m,s,\phi)+NC\ N_b(m) \qquad 4.2.2.1\text{-}1$$

The bandpass filters were all eliptical with a unity gain in the passband (see 4.2.1.2). Table 4.2.2.1-1 gives a summary of the additive colored noise distortions.

4.2.2.2 The Pole Distortion

Figure 4.2.2.2-1 illustrates the implementation of the "pole distortion." The speech is first pre-emphasized using a second order filter, and a framed LPC analysis is performed. The results of the LPC analysis is then used to inverse filter the original speech, giving an approximation of the glottal wave excitation [4.8].

The poles of the vocal tract functions are then found by factoring the LPC polynomial. Then the pole distortion is applied by first identifying all the poles within a fixed frequency range, and then moving them slightly in both frequency and bandwidth. This "jittering" of the poles is controlled by two uniform random number generators. The "frequency range," FR, factor gives the range of frequency, in Hertz, in which the poles are allowed to move. The "bandwidth factor", BF, is a multiplicative factor controlling the bandwidth motion by

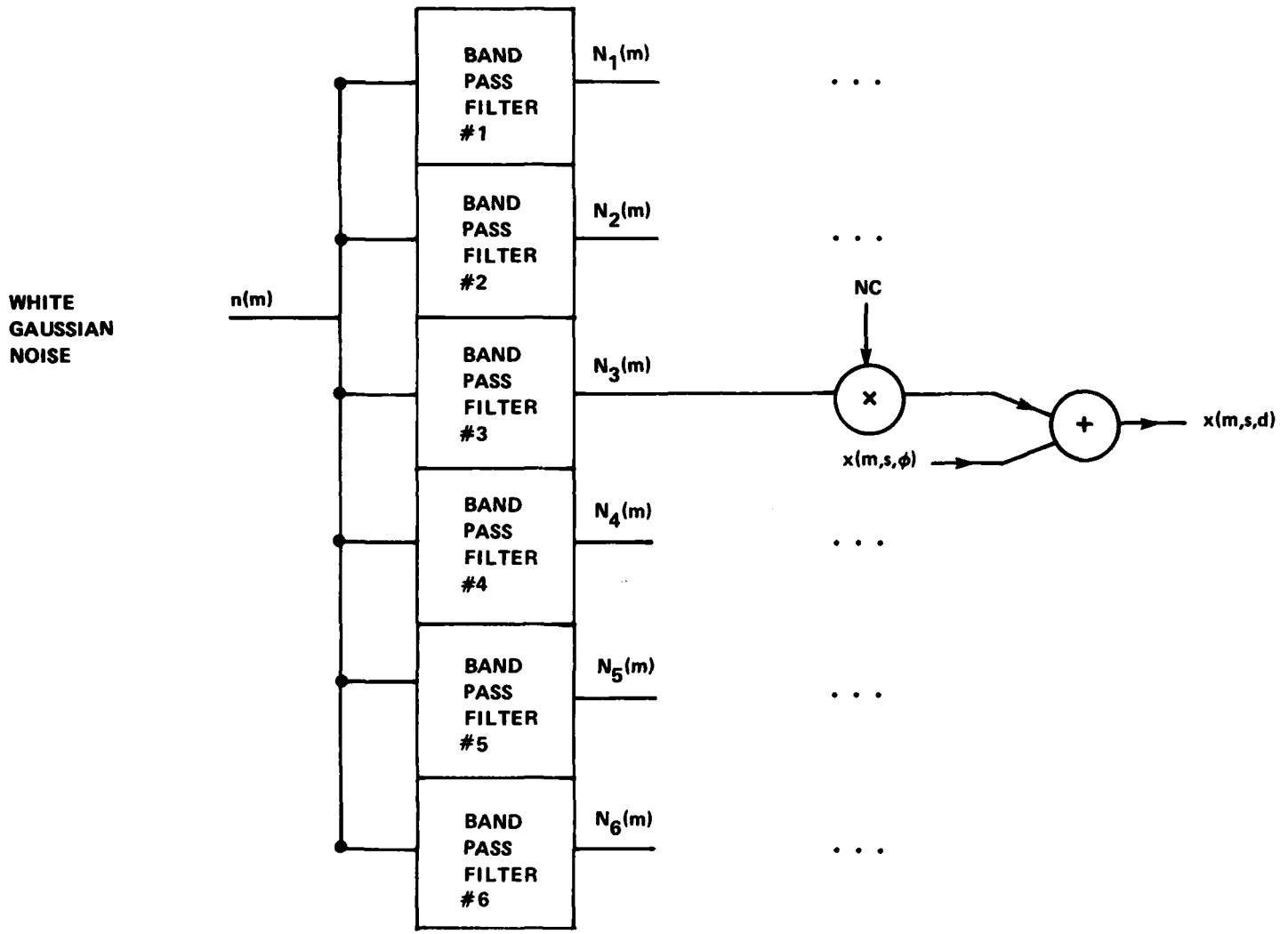


FIGURE 4.2.2.1–1 SYSTEM FOR CREATING THE FREQUENCY VARIANT ADDITIVE NOISE DISTORTION

Bandpass Filter	Noise Constants					
	1	2	3	4	5	6
0-400 HZ	.305	.152	.076	.038	.019	.009
400-800 HZ	.305	.152	.076	.038	.019	.009
800-1300 HZ	.305	.152	.076	.038	.019	.009
1300-1900 HZ	.305	.152	.076	.038	.019	.009
1900-2600 HZ	.305	.142	.076	.038	.019	.009
2600-3400 HZ	.305	.152	.076	.038	.019	.009

Table 4.2.2.1-1. COLORED NOISE DISTORTIONS

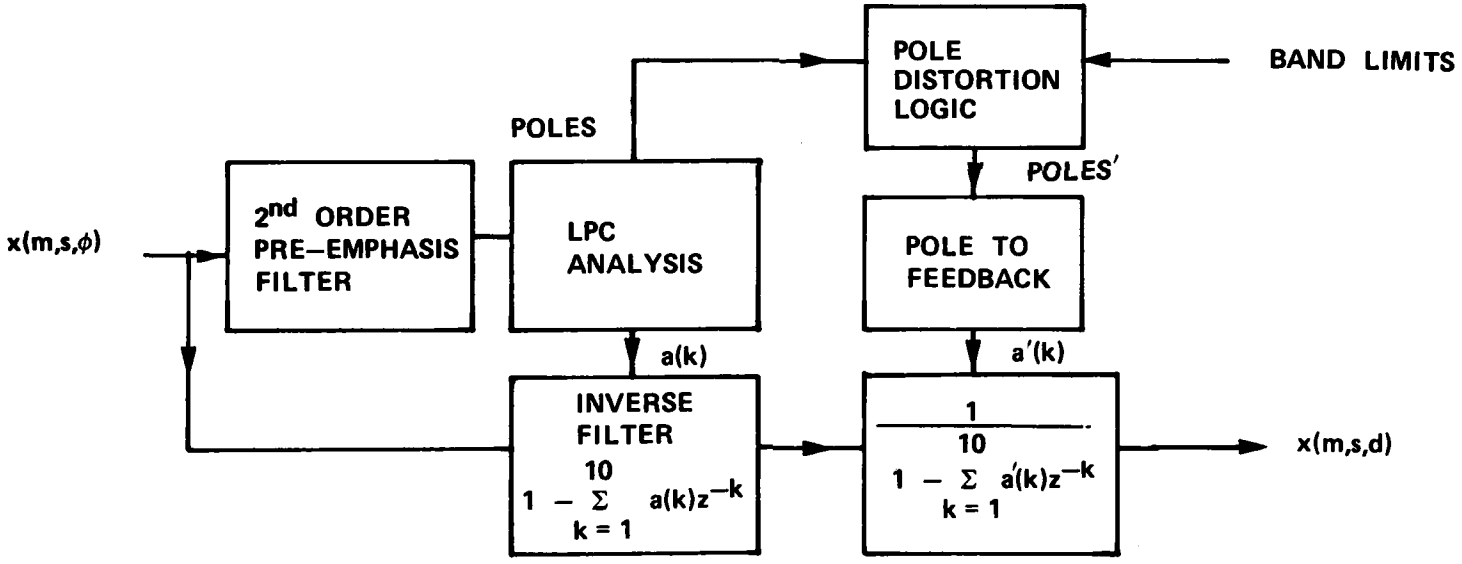


Figure 4.2.2.2–1 System for Producing the Frequency Variant Pole Distortions

distorted radius = (undistorted radius)(1+BF·r) 4.2.2.2-1

where r is a uniform random variable which ranges between plus one and minus one.

Once the pole locations are distorted, they are recombined to form a new set of LPC coefficients, a'(k). These coefficients are used to implement a new vocal tract filter to create the distorted speech.

The pole distortions (PD) are summarized in Table 4.2.2.2-1.

4.2.2.3 The Banded Frequency Distortion

The operation of the banded frequency distortion is illustrated in Figure 4.2.2.3-1. The speech is first windowed using overlapping Hamming windows, where the window length is twice the frame interval, and the frame interval is 128 points. The speech is then transformed using a 256 point FFT. In the frequency domain, noise is then added to the samples in bands. The noise is added with a random magnitude but with a phase equal to the phase of the original speech. Then the samples are inverse transformed back into the time domain and recombined using overlapped adds.

The parameters controlling the banded frequency distortion are the band limits and the standard deviation of the added noise, which is white and Gaussian. Table 4.2.2.3-1 summarizes the banded distortions used in this study.

Pole Distortion
Frequency Distortion

Distortion Band (HZ)	Frequency Range (HZ)					
	1	2	3	4	5	6
200-400	20	40	60	80	100	120
400-800	20	40	60	80	100	120
800-1300	50	90	130	170	210	250
1300-1900	50	90	130	170	210	250
1900-2600	100	150	200	250	300	350
2600-3400	150	200	250	300	350	400

Bandwidth Distortion

Distortion Band	1	2	3	4	5	6
0-400	.025	.05	.075	.1	.2	.3
400-800	.025	.05	.075	.1	.2	.3
800-1300	.025	.05	.075	.1	.2	.3
1300-1900	.025	.05	.075	.1	.2	.3
1900-2600	.025	.05	.075	.1	.2	.3
2600-3400	.025	.05	.075	.1	.2	.3

Table 4.2.2.2-1. POLE DISTORTION CONTROL PARAMETERS

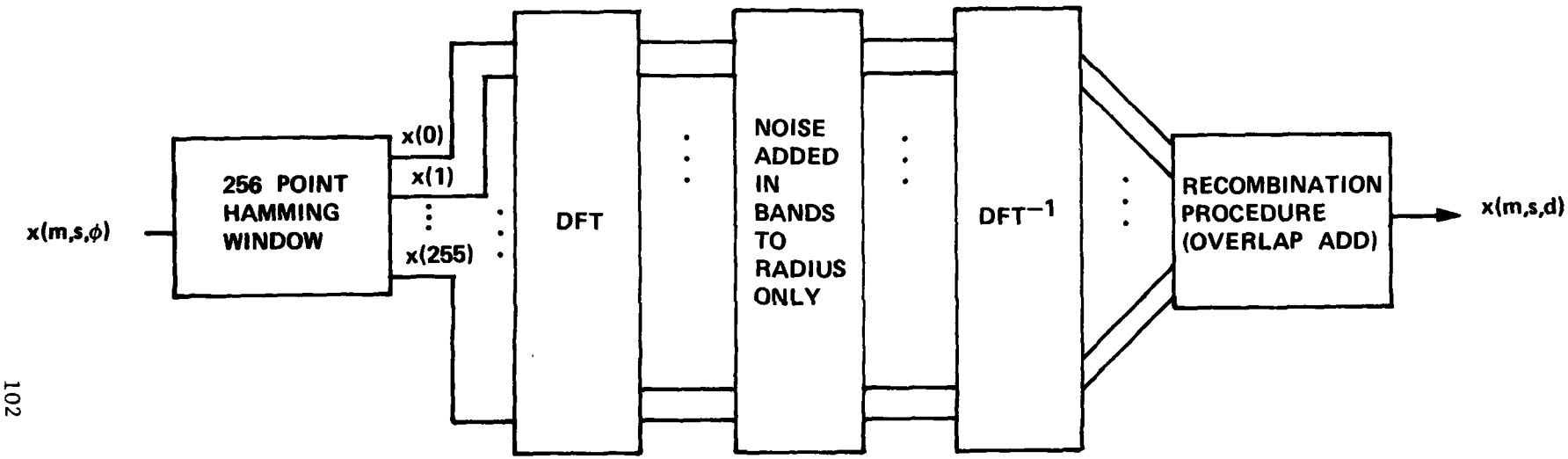


Figure 4.2.2.3–1 System for Implementing the Banded Frequency Distortion

Banded Distortion

Band Limits	Standard Deviation of Noise					
	1	2	3	4	5	6
0-400	.1	.2	.4	.6	.9	1.
400-800	.5	.8	1.1	1.4	1.7	2.0
800-1300	2.0	2.2	2.4	2.6	2.8	3.0
1300-1900	2.0	2.2	2.4	2.6	2.8	3.0
1900-2600	3.5	4.0	4.5	5.0	5.5	6.0
2600-3400	10.	13.	16.	19.	22.	25.

Table 4.2.2.3-1. CONTROL PARAMETERS FOR BANDED NOISE DISTORTION

REFERENCES

4.1. T. P. Barnwell and A. M. Bush, "A Minicomputer Based Digital Signal Processing Laboratory," EASCON '74, Washington, DC, Oct. 1974.

4.2. N. S. Jayant, "Adaptive Delta Modulator with a One Bit Memory," Bell Systems Tech. J., Vol. 49, March 1970.

4.3. A. H. Gray, Jr., and J. D. Markel, "Quantization and Bit Allocation in Speech Processing," IEEE Trans. ASSP, Vol. 24, No. 6, December 1976.

4.4. A. M. Noll, "Cepstrum Pitch Determination," JASA, Vol. 41, No. 2, Feb. 1967.

4.5. A. M. Noll, "Short-Time Spectrum and Cepstrum Techniques for Vocal Pitch Detection," JASA, Vol. 36, No. 2, Feb. 1964.

4.6. R. Zelinski and P Noll, "Adaptive Transform Coding of Speech Signals," IEEE Trans. ASSP, Vol. ASSP-25, No. 4, Aug. 1977.

4.7. J. M. Tribolet and R. E. Crochiere, "An Analysis/Synthesis Framework for Transform Coding of Speech," Proc. ICASSP-79, Washington, DC, April 1979.

4.8. T. P. Barnwell, R. W. Schafer and A. M. Bush, "Tandem Interconnection of LPC and CVSD Digital Speech Coders," Final Report, DCA, DCEC, DCA 160-76-C-0073, November 1977.

CHAPTER 5

EFFECTS OF SELECTED FORMS OF DEGRADATION ON SPEECH ACCEPTABILITY AND ITS PERCEPTUAL CORRELATES

The primary purpose of this phase of the project was to provide criterion measures for evaluating the predictive potential of the various physical voice measures presently under consideration. For this purpose it was essential that representatives of widely diverse forms of degradation be included among the conditions evaluated. Among the forms considered are those inherent in the simplest types of analog speech transmission as well as those associated with the most elegant digital voice coding and transmission techniques in use today. Only with such diversity could any assurance be had that observed correlations between specific physical voice measurements and various subjective criteria will obtain for more than a narrow class of distortions.

A second purpose to be served by this phase of the project was the cross validation of the DAM itself. Since the DAM was developed as the result of a comprehensive examination of the effects of representative types of degradation (including many of those treated in the present investigation) on various subjective criteria of acceptability, the results of the present investigation permit a rigorous test and possible refinement of DAM administration and scoring procedures.

Finally, depending on the configurations of DAM scores produced by various novel forms of degradation, some insights may be gained which permit improvements in current technology of acceptability prediction from physical voice measurements. Conceivably, novel techniques may also be suggested by these results in combination with results bearing on the

efficiency of specific prediction techniques for specific classes of degradation.

All of the above purposes are given consideration as appropriate in the course of discussing the results presented in the following sections.

5.A Methods and Materials

5.A.A Listening Crews

Professional listening crews (young adults of both sexes) of eight to ten members participated in all evaluation sessions conducted under the project. On the basis of a retrospective criterion of self-consistency within each testing session, one or more members were eliminataed such that the data for the eight most self-consistent members were retained for analysis.

5.A.B Speakers

Four speakers, three males and one female, were used for all evaluations. The ordering of experimental system-conditions varied from one speaker to the next in a systematic manner designed to minimize time-order effects on the data for any system-condition. Twenty-four system-conditions, two anchors and four probes were evaluated in each testing session. The anchors and probes were always presented at the beginning of each series of system-conditions involving a given speaker. The ordering of anchors and probes was randomly determined in each instance. Whenever possible, several distinct types of degradation were represented in a given session. System-conditions were effectively randomly ordered within a session for one speaker, and then systematically reordered for the remaining speakers to provide some amount of counterbalancing and, thus, to soften the effects of any inter-condition influences.

5.B Experimental Results

Presented in the following sections are DAM score patterns for the various forms of degradation. For each class of degradation the diagnostic patterns are presented in separate sub-figures for male (average of three) and female speakers. Except where pronounced sex differences are evident, the discussion will be addressed primarily to the results for the male speakers. Primary interest in these figures attaches to the Composite Acceptability score (C-A) and the parametric score for acceptability, (P-A), intelligibility (P-I), and pleasantness (P-P). Although it is one of the components of C-A, the isometric acceptability score (I-A) is not included in the graphic portrayals. The reason for this is that it has a virtually perfect correlation (.994) with the average of the parametric intelligibility and parametric pleasantness scores. Of considerable but secondary interest are the "diagnostic patterns" of perceptual quality scores. Depending on the form of degradation involved, diagnostic score patterns for experimental systems may provide insights of substantial value for purposes of remedial action. Here, they serve primarily to enhance our basic knowledge of the perceptual affective consequences of speech degradation and to reveal further useful features of the DAM.

Two administrations of the DAM, separated by intervals of four to six weeks, were performed for all the system-conditions except those involving pole distortions and band distortions. With exception of these later cases, all results presented in the following sections are thus averages based on response data from two administrations.

5.1 Degradation by Coding

Treated in this section are cases of distortion which are intrinsic

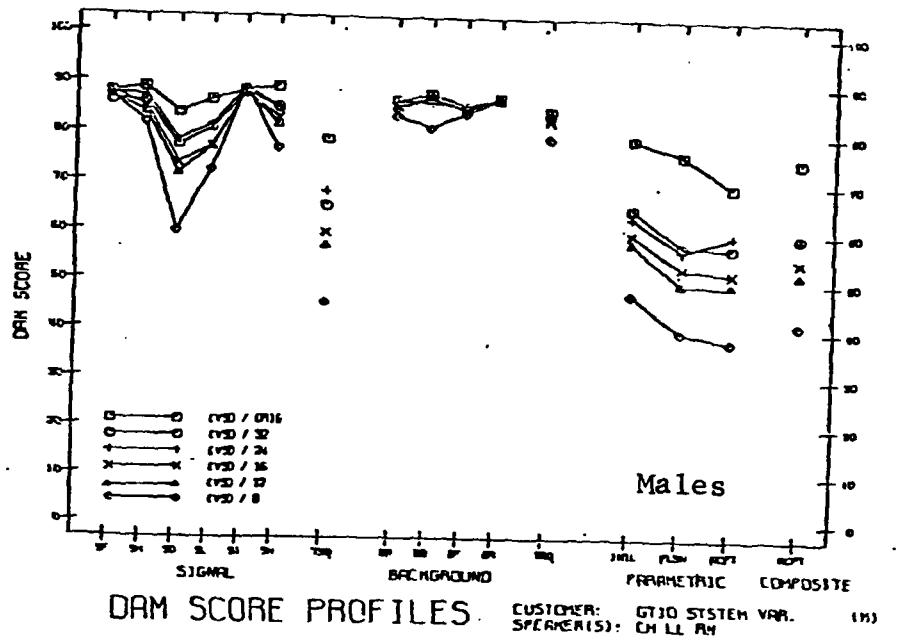
to various speech coding techniques and, in a sense, reflect the inadequacies of such techniques. In one category are various broadband wave form-preserving techniques in which a major source of degradation is quantization of the speech signal. In the second category are various, more complex predictive coders.

5.1.1 Simple wave-form coders

The wave-form coders treated in this investigation are CVSD, ADM, APCM, and ADPCM which are described in section 4.1.1.

5.1.1.1 Effects of continuously variable-slope delta modulation (CVSD) on DAM scores

Five realizations of CVSD technique, which differed only with respect to data rate, were treated in this investigation. A control condition involving essentially unprocessed speech was included within the same DAM testing session. The DAM results are presented in Fig. 5.1.1.1. In all major respects they are typical of previous DAM results for CVSD [5.1]. Except in the case of the lowest data rate, background quality is negligibly affected by CVSD. Listeners evidently do not confuse quantization "noise" with true noise. Rather, they correctly perceive it as distortion: the SD scale of the DAM is the most sensitive of the perceptual quality measures. The present results differ somewhat from those of previous studies in that they show consistent, though not pronounced reductions in scores on the SH, SL and SN scales as data rate is reduced. Such results are most typical of conditions involving audio pass-band restriction and may, therefore, have a rational basis in the present case.



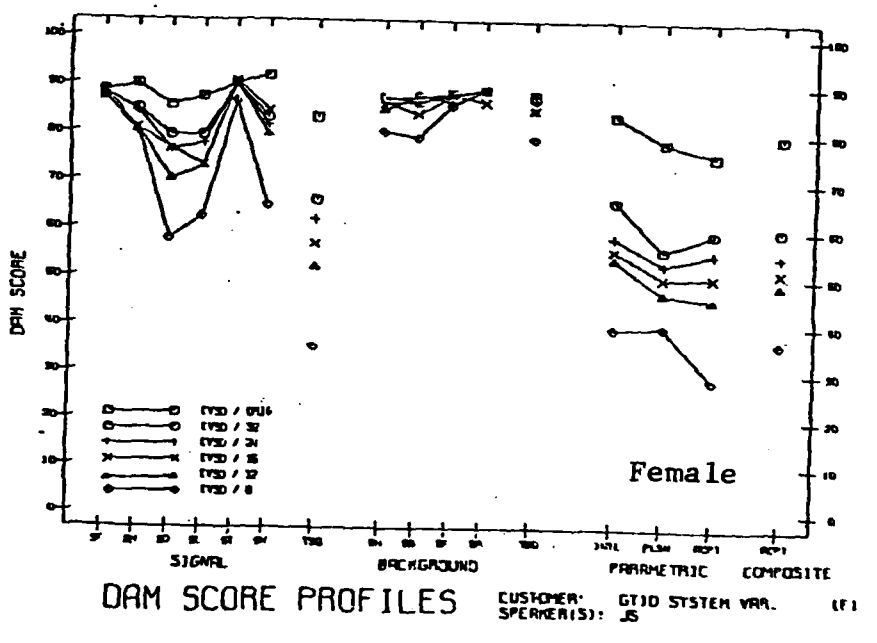


Figure 5.1.1.1 Effects of continuously-variable slope delta modulation on DAM scores for male and female speakers.

For reasons that are not immediately obvious, scores for the control case and for the case of 32K bps CVSD are generally somewhat lower than those previously obtained for nominally comparable conditions, though scores for the 16K bps case are very close to historical norms for this condition.

5.1.1.2 Effects of adaptive delta modulation (ADM) on DAM scores

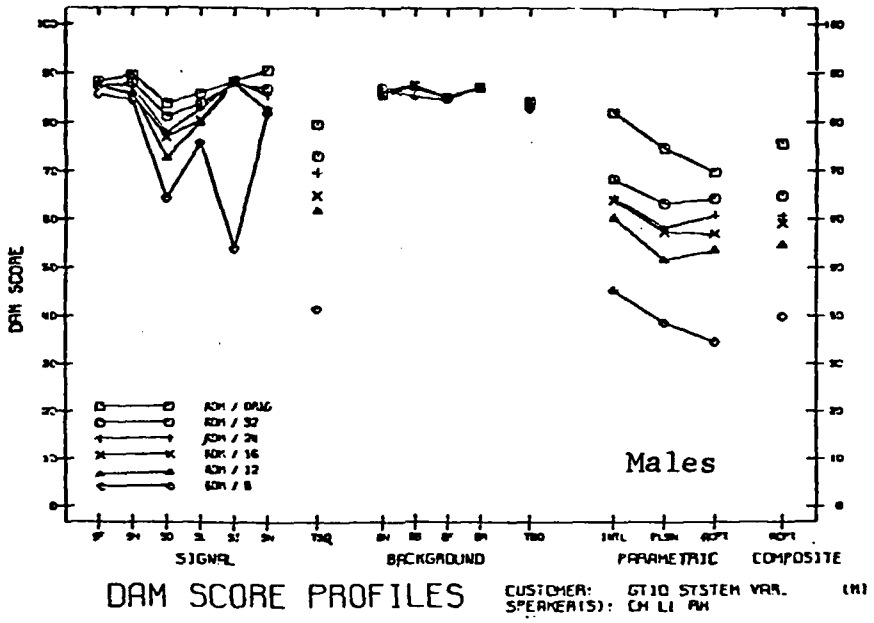
Figure 5.1.1.2 presents DAM results for the case of adaptive delta modulation. Predictably, perhaps, they are quite similar to those for CVSD at most corresponding data rates. An exception is the case of ADM at 8K bps where a severely depressed score on the SI (signal interrupted) scale can be observed. It is quite possible that this result can be attributed to an experimental artifact, but further investigation will be needed to resolve this issue. This is not of great interest since no one has seriously suggested using such a coding procedure at this rate.

5.1.1.3 Effects of adaptive pulse code modulation (APCM) on DAM scores

Figure 5.1.1.3 presents DAM results for the case of adaptive pulse code modulation techniques. As in the two previous cases the subjective consequences of this type of coding are confined almost exclusively to signal quality. Here their general form is quite similar to those for the cases of CVSD and ADM but for a small, though consistent, depression of scores on the SI scale. However, the general level of scores for APCM is substantially lower than for CVSD and ADM.

5.1.1.4 Effects of adaptive differential pulse (ADPCM) code modulation on DAM scores

Figure 5.1.1.4 shows DAM scores for adaptive differential pulse code modulation. The results for this condition are quite similar to those



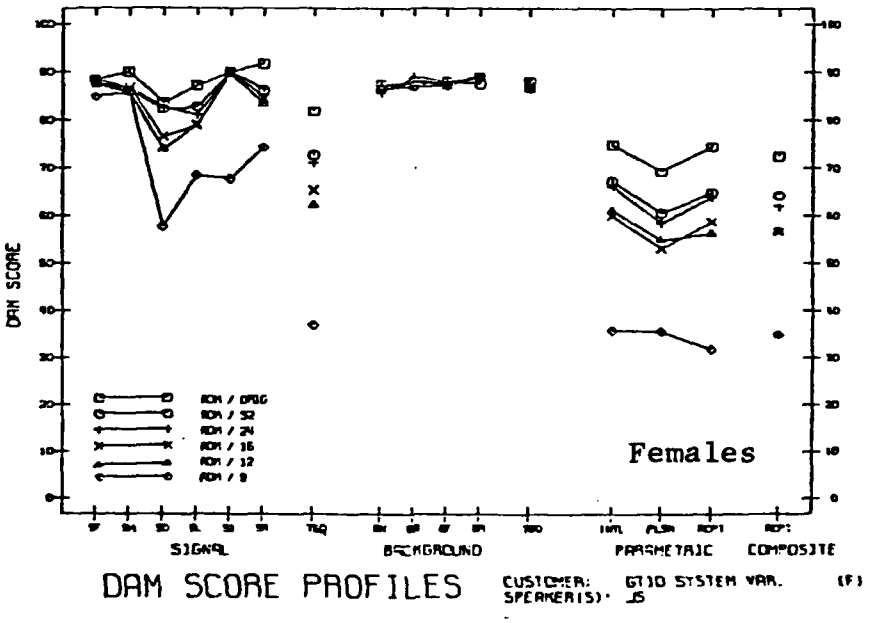
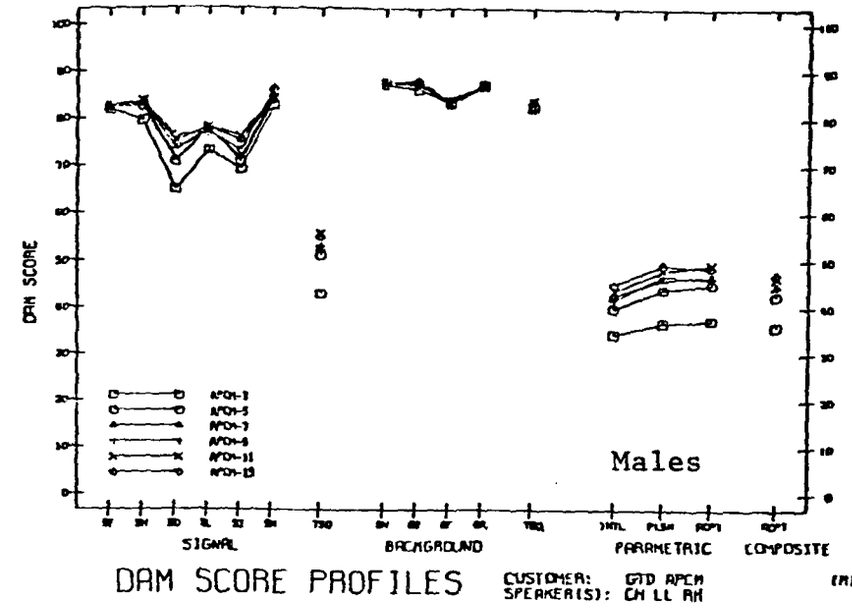


Figure 5.1.1.2 Effects of adaptive delta modulation on DAM scores for male and female speakers.



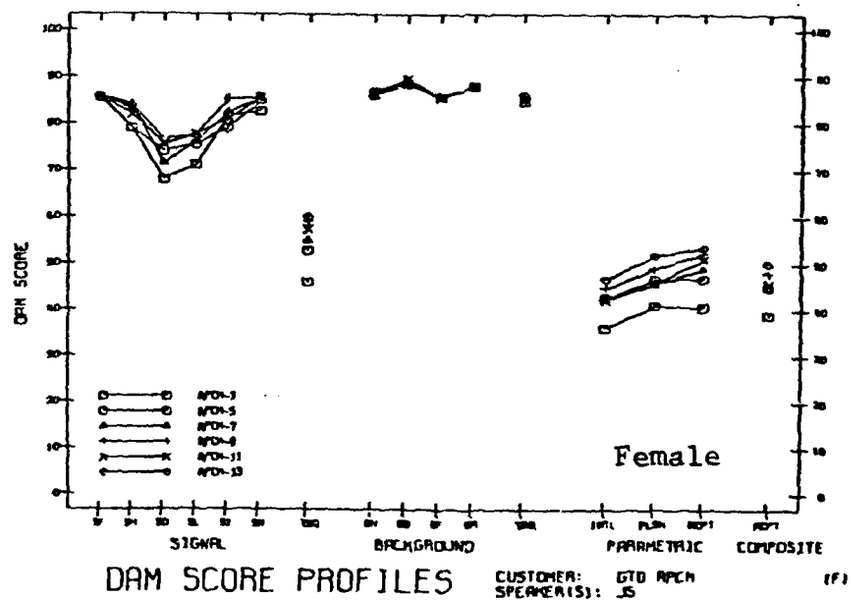
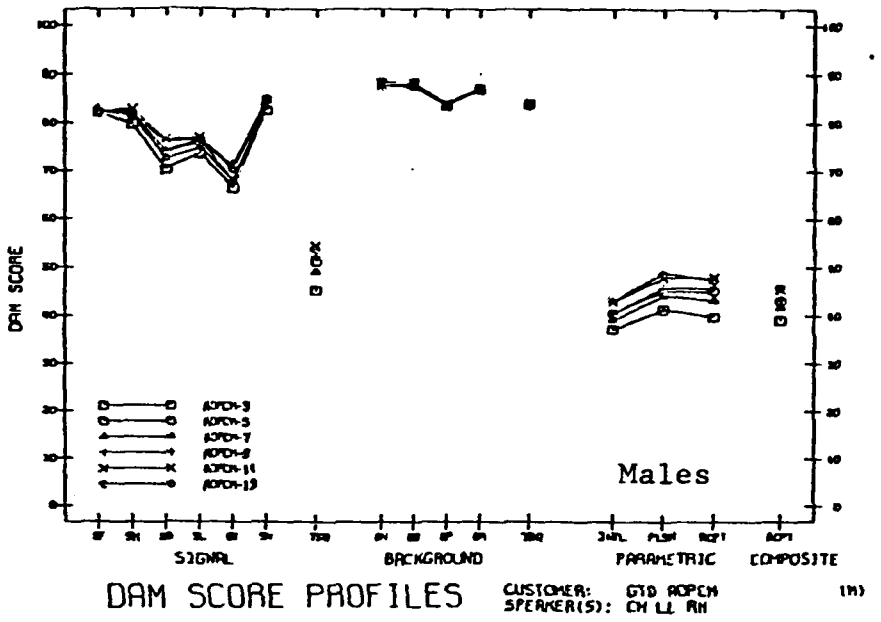


Figure 5.1.1.3 Effects of adaptive differential pulse code modulation DAM scores for male and female speakers.



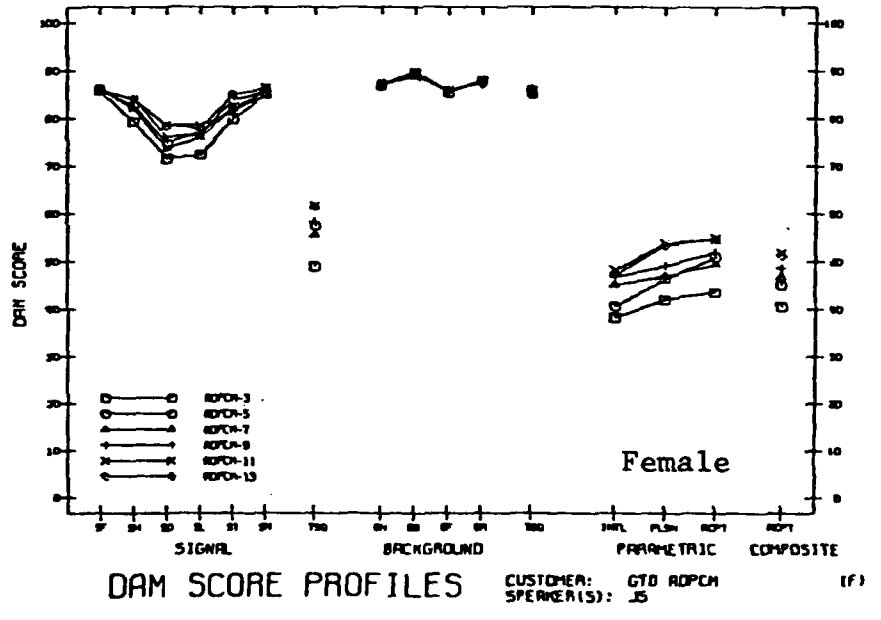


Figure 5.1.1.4 Effects of adaptive differential pulse code modulation on DAM scores for male and female speakers.

for APCM. It would appear that adaptive differential pulse code modulation does not significantly improve acceptability over APCM at comparable degrees of quantization, but may, however, at comparable transmission data rates with optimal channel coding. Qualitatively, ADPCM would appear to sound somewhat less distorted but more interrupted than APCM.

5.1.2 The effects of linear predictive coding on DAM scores

The linear predictive coder used in this investigation is described in Section 4.1.2. Figure 5.1.2 shows that the present realization of LPC in the range of 2-2.9K bps yields DAM score patterns and overall levels very similar to those of the normative 2.4K bps LPC as reported by Voiers [5.1]. Normative DAM results for the higher data rates are not available for systems without error correction, so that comparisons are not possible for the higher data rates. However, the present results indicate that increasing the data rate to 3.9K bps significantly improves the quality of LPC-processed speech, though further increases do not appear to be beneficial. On the other hand, it appears that digitization at high data rates does not significantly impair quality obtained with analog LPC techniques. (LPC/Orig. in Figure 5.1.2)

5.1.3 The effects of adaptive predictive coding on DAM scores

Figure 5.1.3 shows the effects of APC on DAM scores. The parameter in these graphs is bits/frame which is associated with data rates of from 13333 to 15867 bps.

Perceptual quality score patterns for APC are quite similar to those for LPC. Though score levels are generally somewhat higher for APC, this superiority is evidently achieved only at enormous cost in transmission data rate. Listeners perceive significant amounts of signal and

background "flutter" (SI scale) and raspiness (SD scale).

5.1.4 Effects of voice-excited vocoding (VEV) on DAM scores

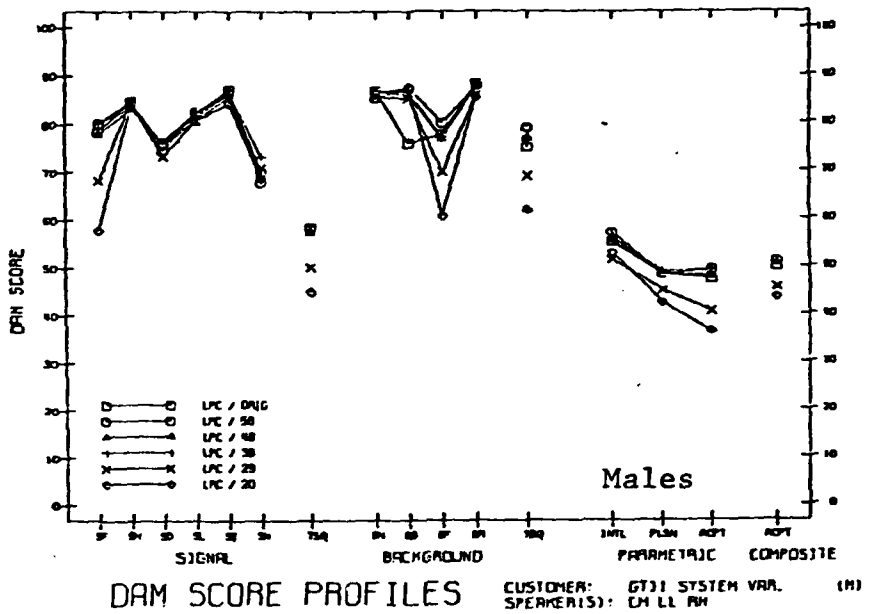
The voice excited vocoding technique, described in Section 4.1.4 is essentially a modification of the APC technique treated above. Two realizations of VEV were examined here. In the first (Fig. 5.1.4.1) the voice band had seven level quantization; in the second (Fig. 5.1.4.2), thirteen level quantization. The parameter in each case is PARCOR frame rate. Differences between the subjective effects of these two techniques are very small. Listeners possibly perceive a slightly greater degree of signal and background flutter with coarser quantization in the case of male speakers, but this trend is absent in the case of the female speaker. Generally, more background flutter is perceived in the case of VEV than in the case of APC.

5.2 Controlled Degradation

Treated in this section are various basic types of speech degradation, one or several of which may be encountered in most speech-communication situations. They are distinguished from the coding distortions dealt with in Section 4.1 by the fact that they are generally not deliberately introduced but occur rather as by products of various coding techniques or channel characteristics.

5.2.1 Simple forms of controlled degradation

Seven of the commonly encountered forms of degradation are dealt with in the following sections. They include broad band-limited Gaussian noise, frequency passband restriction, interruption, peak and center clipping, coarse quantization and echo.



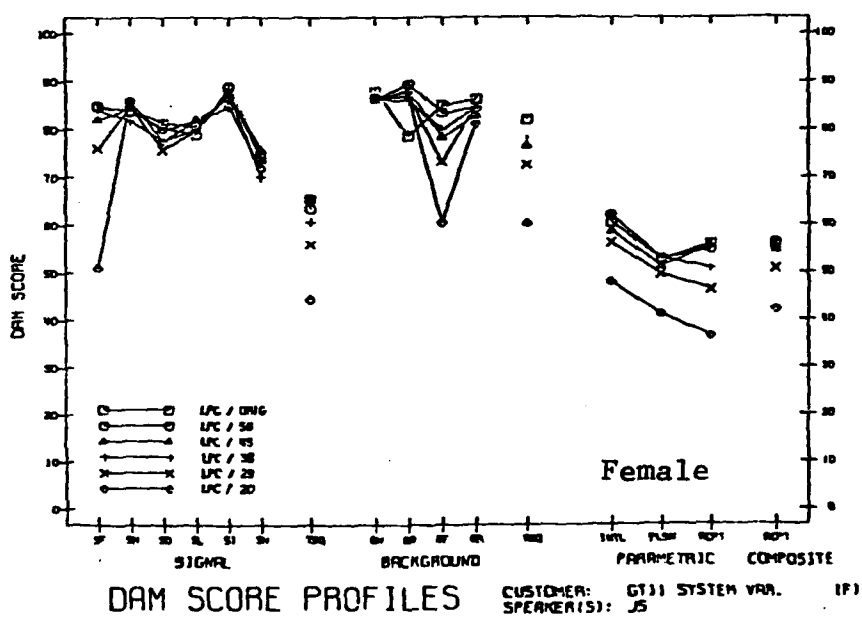
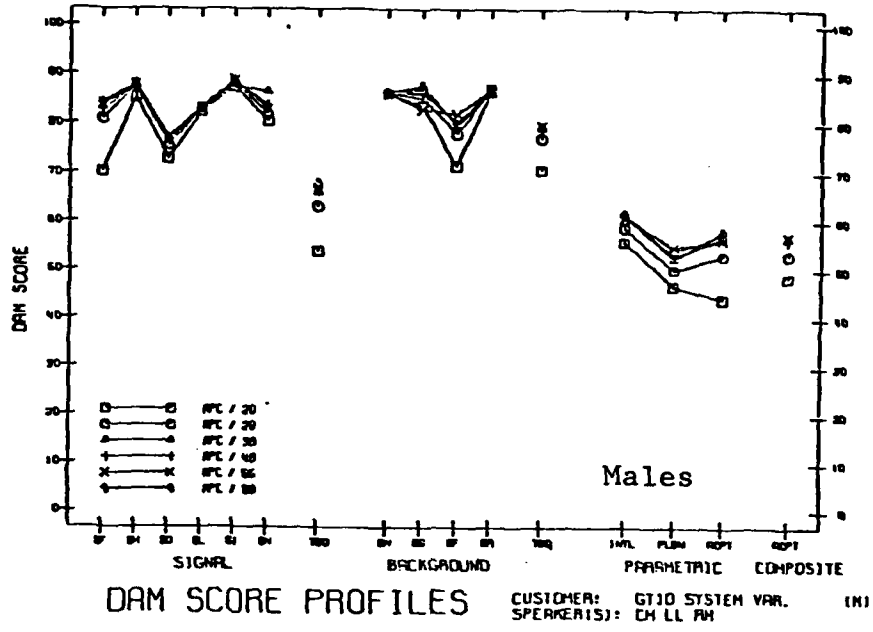


Figure 5.1.2 Effects of linear predictive coding on DAM scores for male and female speakers.



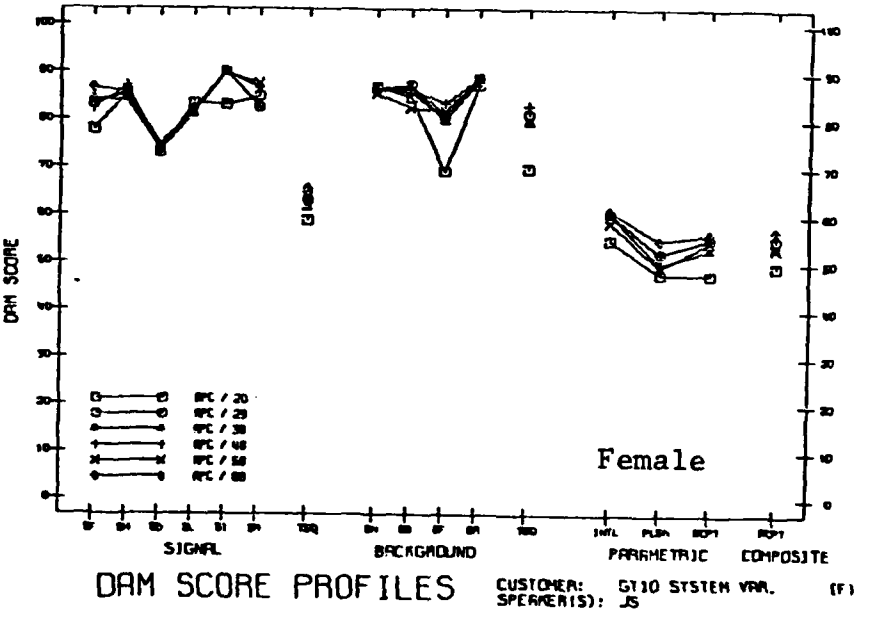
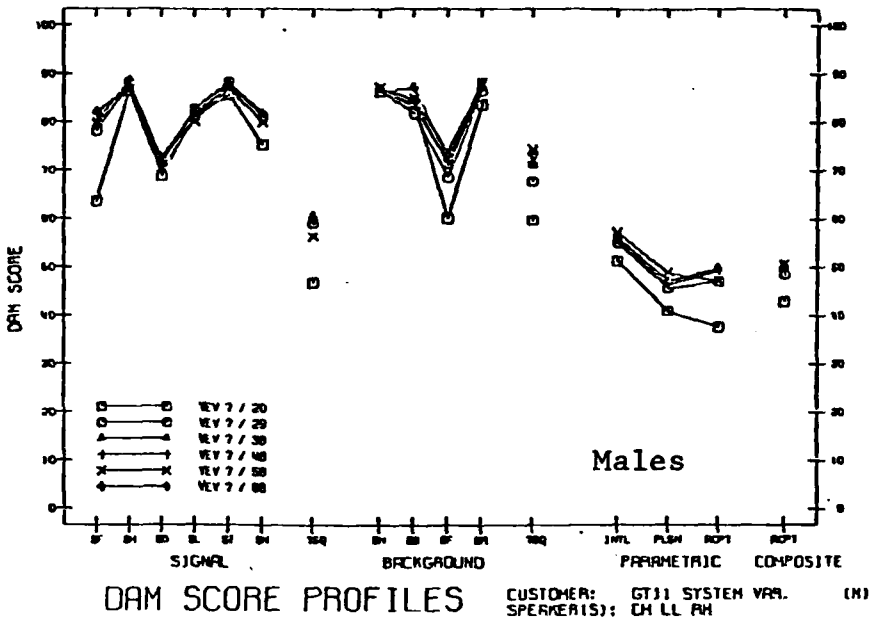


Figure 5.1.3 Effects of adaptive predictive coding on DAM scores for male and female speakers.



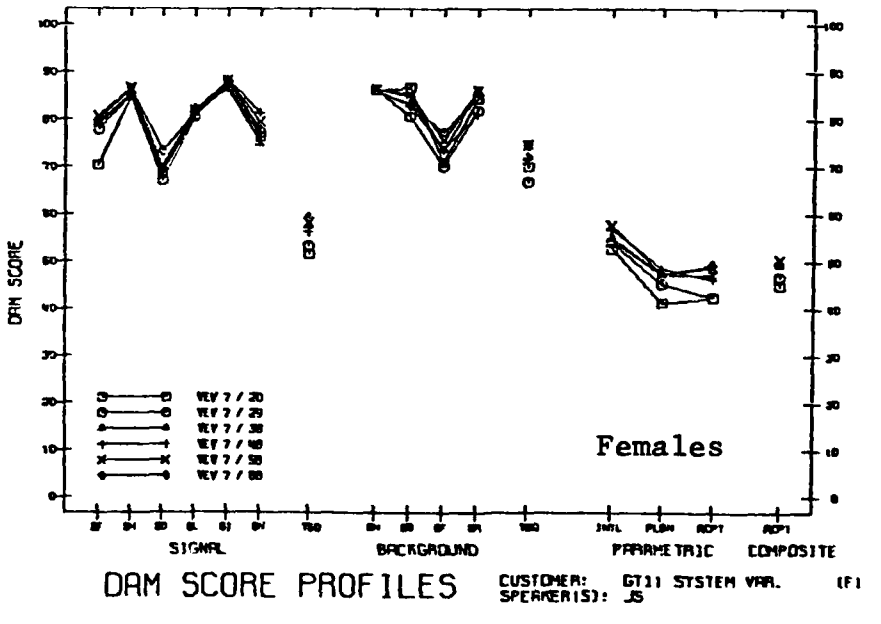
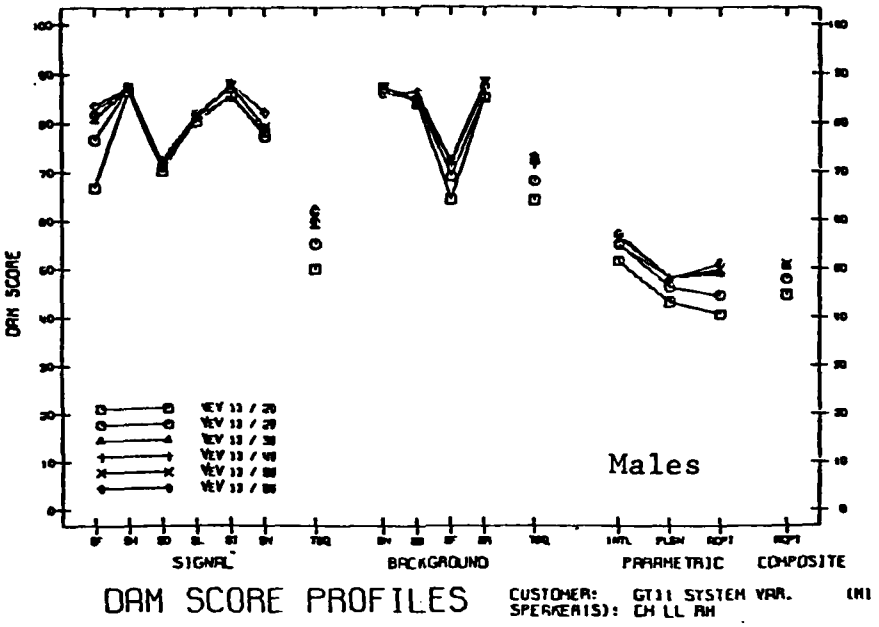


Figure 5.1.4.1 Effects of Voice-excited Vocoding (7 level quantization) on DAM scores for male and female speakers.



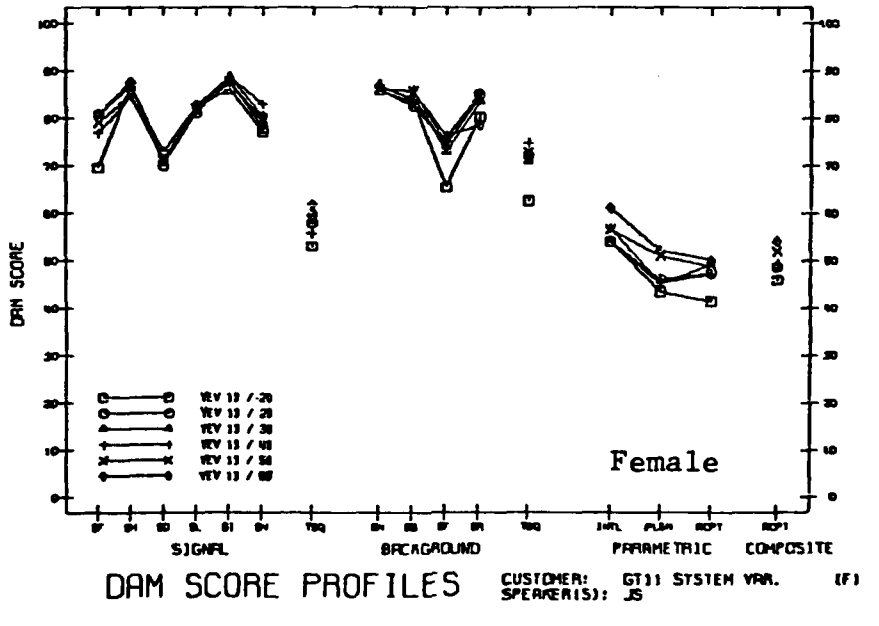
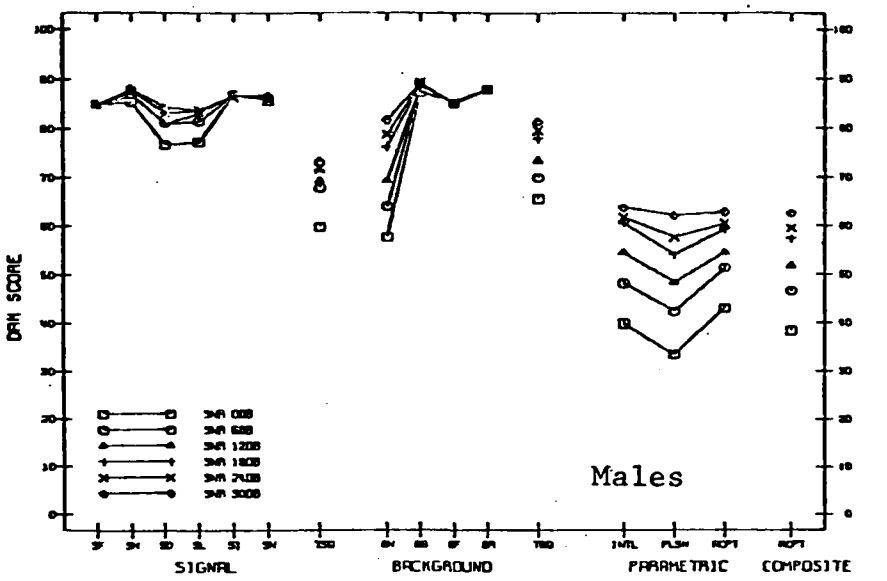


Figure 5.1.4.2 Effects of Voice-excited Vocoding (13 level quantization) on DAM scores for male and female speakers.

5.2.1.1 Effects of additive broad-band noise on DAM scores

Noise is the most umbiquitous of all deterrents to efficient communications. Accordingly, its effects on DAM scores merit special consideration.

Figure 5.2.1.1 shows DAM diagnostic profiles for six conditions of S/N ratio (4K Hz passband for the speech and noise). From the figure it is clear that the SN scale is the most sensitive to additive Gaussian noise, but the results again illustrate an important principle of the psychophysics of speech: In virtually no instances are the consequences of degradation with respect to a single stimulus parameter confined to a single stimulus elementary phychological parameter. It has long been known, for example, that whereas the elementary psychological parameter, pitch, depends primarily on stimulus frequency, it also varies with stimulus intensity, duration, and complexity. In the present case, the mechanism whereby values on the SL scale (normally most sensitive to high frequency attenuation) also decreases with S/N ratio is easily specified: The spectrum level of typical speech is highest in the region of 500 Hz but decreases at approximately 9 dB per octave both above and below that region. A secondary effect of uniform spectrum noise, therefore, is generally that of passband restriction, particularly at the upper end of the speech spectrum. Less readily predicted, but by no means contraintuitive, is a slight reduction of the SD scale (the scale most sensitive to amplitude distortion). With extremely unfavorable S/N ratios, listeners are evidently not able to make the noise vs. distortion distinction with the same ease that they accomplish this under less severe conditions of degradation.



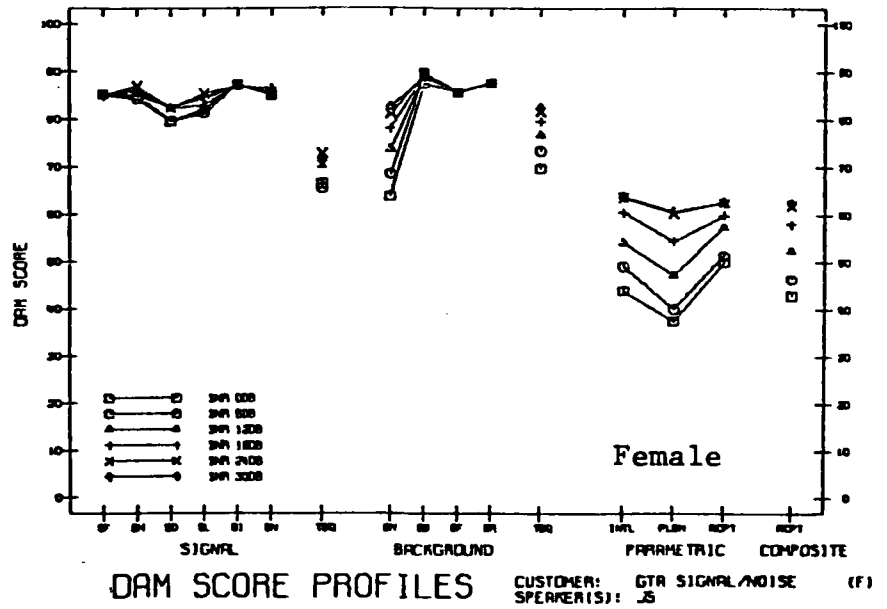


Figure 5.2.1.1 Effects of broad-band Gaussian noise on DAM scores for male and female speakers.

5.2.1.2 Effects of frequency on DAM scores

The results of research leading to the development of the DAM served in many instances to confirm the principle that even the simplest forms of signal impoverishment have relatively complex subjective consequences: Although the effects of a given form of distortion may be most pronounced in one subjective dimension, they are usually evident in two or more dimensions. On the other hand, many forms of degradation may have a common perceptual effect, as well as unique perceptual consequences. In fact, perceptual quality scales SH, SL, and SN were found to be sensitive in varying degrees to the effects of three major forms of frequency distortion.

All forms of passband restriction have previously been observed to affect the SL scale, which is associated with the perceptual qualities of muffledness, dullness, etc. The effects of high frequency attenuation were found to be confined primarily to this DAM parameter, though some depression of scores on the SH scale was observed with extremely severe high frequency attenuation.

The effects of low frequency attenuation, i.e., high-pass filtering, were observed to be most pronounced in the case of the SH scale, which was in fact designed primarily to sense such effects. However, the SL scale was also observed to be sensitive to high-pass filtering in lesser degree. A third scale, (SN) which is concerned with the perceptual quality of nasality, was found to be sensitive to passband width restriction more or less without regard to the location of the passband. Present results will be seen to provide strong confirmation of findings of the earlier validation studies of the DAM, though sharper filtering was achieved here than previously. In the present investigation frequency filtering was

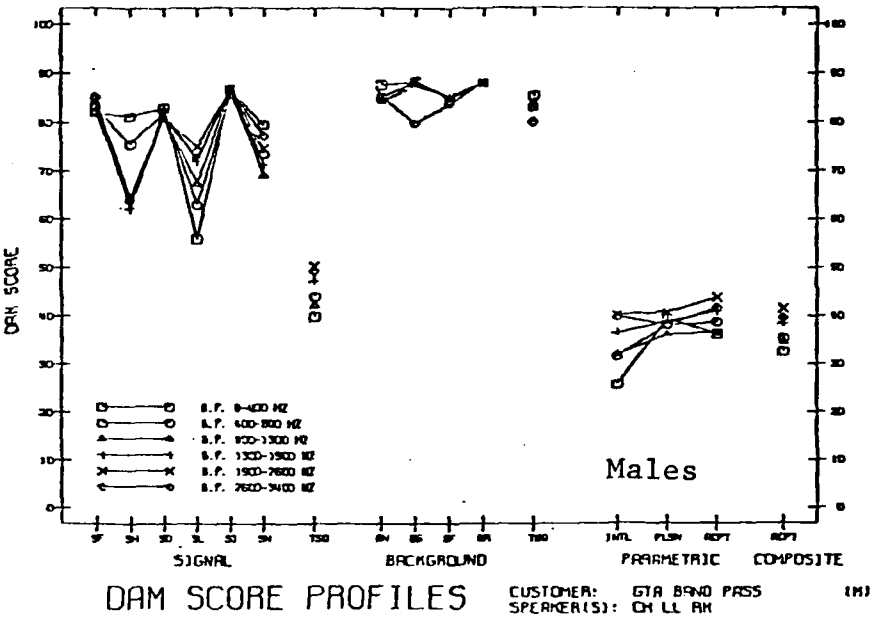
achieved by means of elliptic digital filters with 40 dB or less ripple and transition bands equal to 5 percent of the nominal cutoff frequencies.

5.2.1.2-1 Effects of bandpass filering on DAM scores

Figure 5.2.1.2-1 shows the effects of bandpass filtering on DAM diagnostic score patterns. Consistent with previous findings, three of the primary perceptual quality scales (SH, SL, and SN) are sensitive, but in different ways, to this form of signal distortion. Differences between the trends of the SH scores and SL scores are best rationalized in terms of the character of the rejected band(s) associated with each condition. To the extent that high frequency rejection predominates, SL scores suffer greatest reduction, whereas SH scores reflect the predominance of low-frequency rejection. Scores on the SN scale vary in a more complex manner with the location of the passband, being highest for the high and low extremes, lowest for those passbands near the middle of the frequency scale, in particular those covering the frequency range of the second format. Generally, the scales pertaining to background qualities are little affected by passband restriction. The one case in which a background exhibits depression (BB scale in the case of the 2600-3400 Hz condition) is quite possibly due to an increase in hum associated with the higher gains needed for the high frequency bands.

5.2.1.2-2 Effects of low-pass filtering on DAM scores

From Figure 5.2.1.2-2 it is evident that the effects of low-pass filtering are confined primarily to the SL scale, a result consistent with the purpose for which this scale was designed. Although some variation in other signal quality scales is evident, no consistent trends emerge. All of the background quality scales are virtually "blind" to this form of



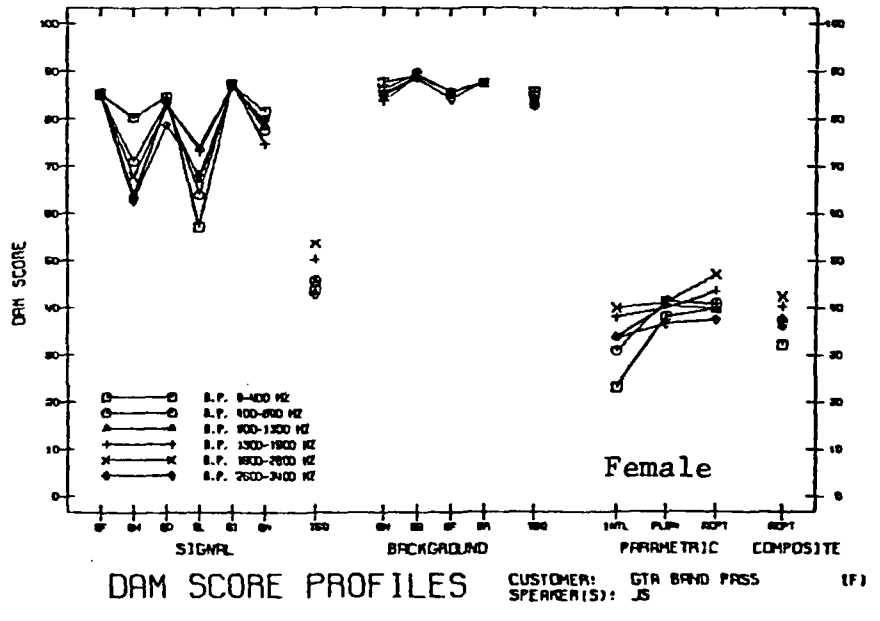
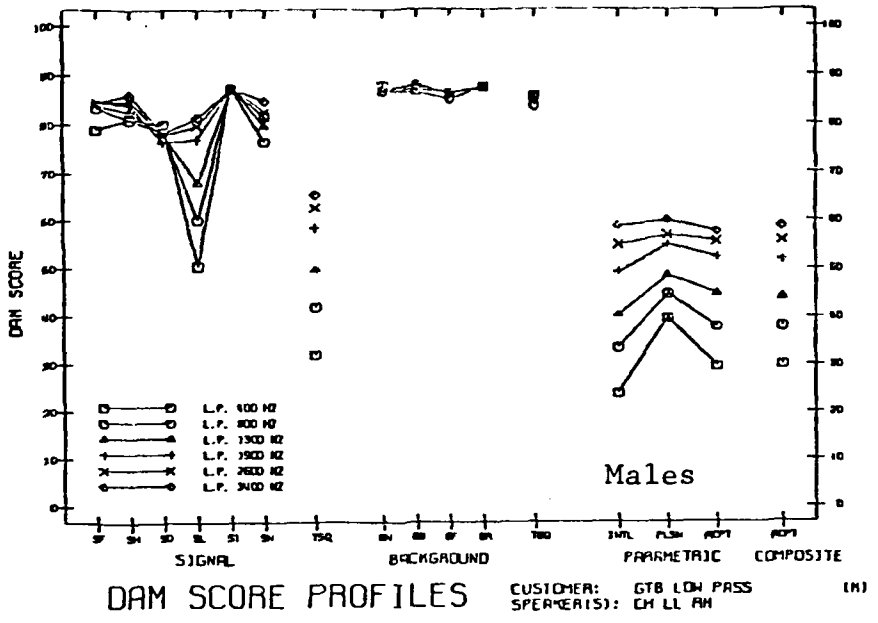


Figure 5.2.1.2-1 Effects of band-pass filtering on DAM scores for male and female speakers.



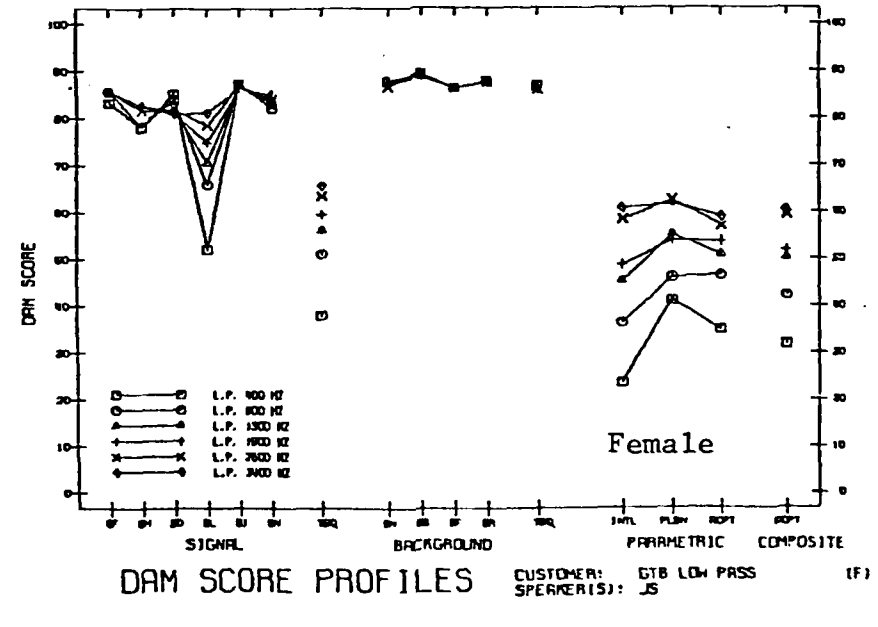


Figure 5.2.1.2-2 Effects of low-pass filtering on DAM scores for male and female speakers.

degradation. It may be of some interest to note that neither the most sensitive perceptual quality nor overall acceptability are substantially affected until attenuation of frequencies below 2K Hz occurs. The fact that the parametric scale for intelligibility is more uniformly affected by high frequency attenuation is consistent with results on the effects of high frequency attentuation on objectively measured intelligibility. It is perhaps consistent with common intuition that parametric pleasantness is generally the least affected of the higher order qualities.

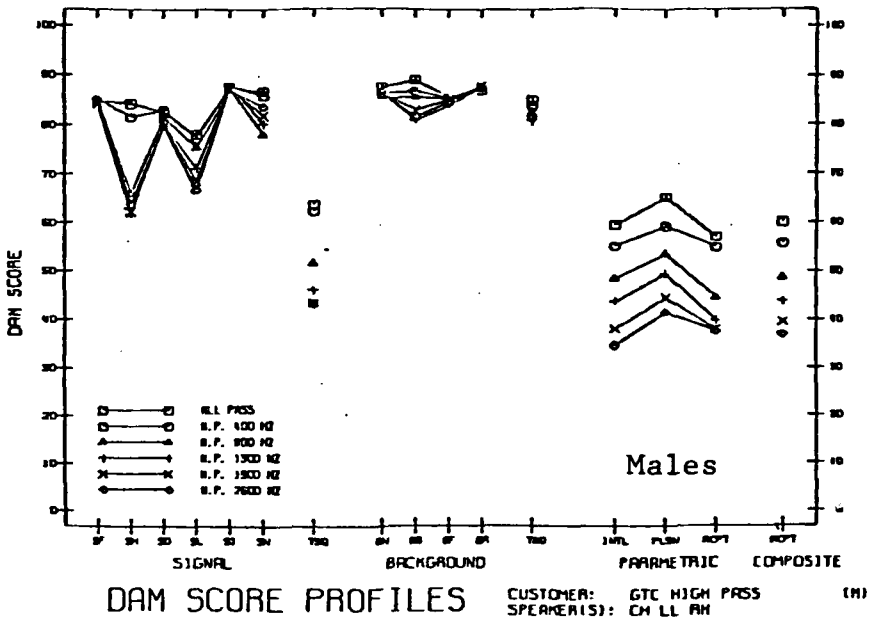
5.2.1.2-3 Effects of high-pass filtering on DAM scores

Four perceptual quality scales appear sensitive to high-pass filtering as shown in Figure 5.2.1.2-3. As expected, the SH scale ultimately exhibits the greatest depression, but two other scales, SL and SN, are more sensitive to moderate degrees of high-passs filtering. Only after frequencies as high as 800 Hz are attenuated does the signal appear to acquire the characteristic "high-pass quality."

Again, a decrease in BB scores with increased high-pass filtering is possibly an experimental artifact. The fact that no comparable trend in BB scores is evident in the case of the female speaker adds credibility to such an explanation.

5.2.1.3 Effects of periodic interruption on DAM scores

Two interruption rates, each with six signal-duty factors were treated in this investigation. In the first case, the signal was interrupted once every 300 samples or 26.6 times a second. The duration of each interruption was then varied from 10 to 150 samples, i.e., from 1.25 milliseconds to 18.75 milliseconds. In the second case, the signal was



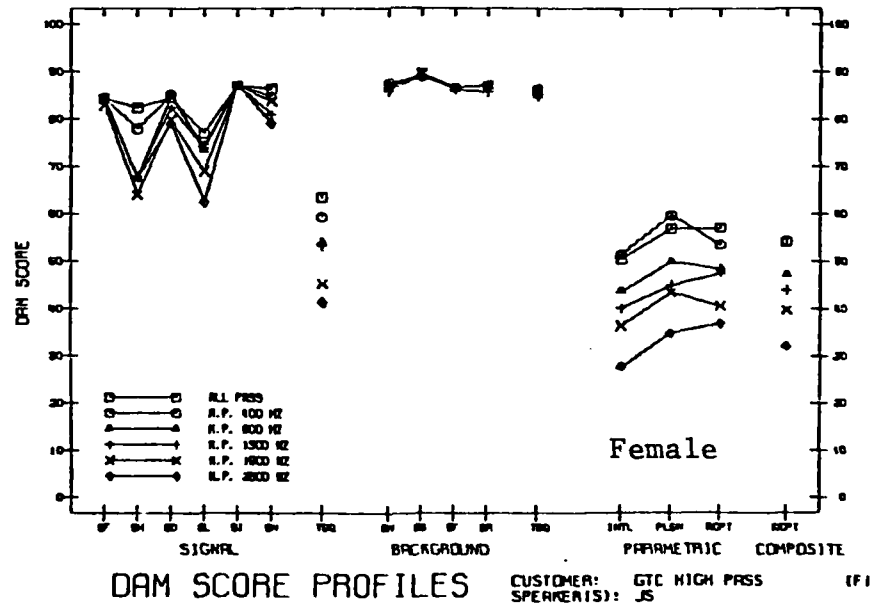


Figure 5.2.1.2-3 Effects of high-pass filtering on DAM scores for male and female speakers.

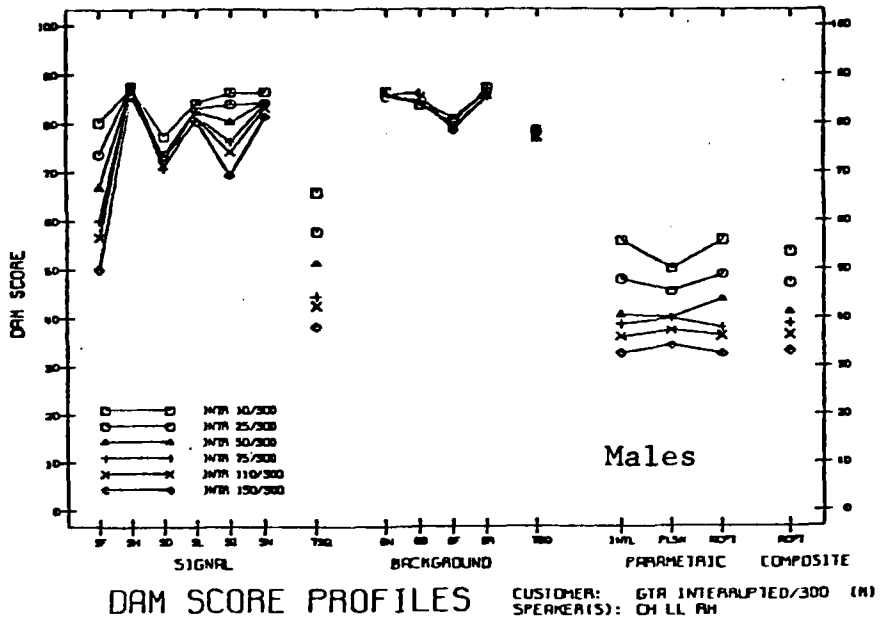
interrupted once every 1024 samples or 7.8125 times a second. Duration of these interruptions was varied from 2 milliseconds to 64 milliseconds.

Figure 5.2.1.3-1 shows that the predominant perceptual quality associated with the more rapid interruption rate is "signal flutter," which quality becomes increasingly pronounced as signal duty factor decreases. Listeners perceive the signal to be interrupted with increasing duration of interruption, but the interrupted quality is less salient than the fluttering of quality. Figure 5.2.1.3-2 shows the effects of less frequent interruption. The fluttering quality is still pronounced but the interruption is now more apparent and in fact predominates in the case of the lowest signal duty factor (.50). Moreover, listeners appeared less inclined in this case to perceive the background as fluttering than they did in the case of more rapidly interrupted speech.

5.2.1.4 Effects of Peak Clipping on DAM scores

Two forms of amplitude distortion are potentially present in many voice communications channels. They can be simply described as "peak clipping" (clamping) and "center clipping," (as might result from intermodulation distortion). It was out of concern for the first of these that the SD scale of the DAM was developed. However, no special provision for the latter was made in the design of the DAM.

Figure 5.2.1.4 shows the effects of six levels of peak clipping on DAM score patterns. The values associated with each condition indicate levels at which peak clipping occurred on a scale where the rms amplitude of the unclipped speech signal was approximately 1350. Two perceptual quality scales SD and BB appear sensitive to this form of degradation. However, an experimental artifact is possibly involved in the case of the latter scale. As noted earlier, the BB scale is quite sensitive to 60 Hz



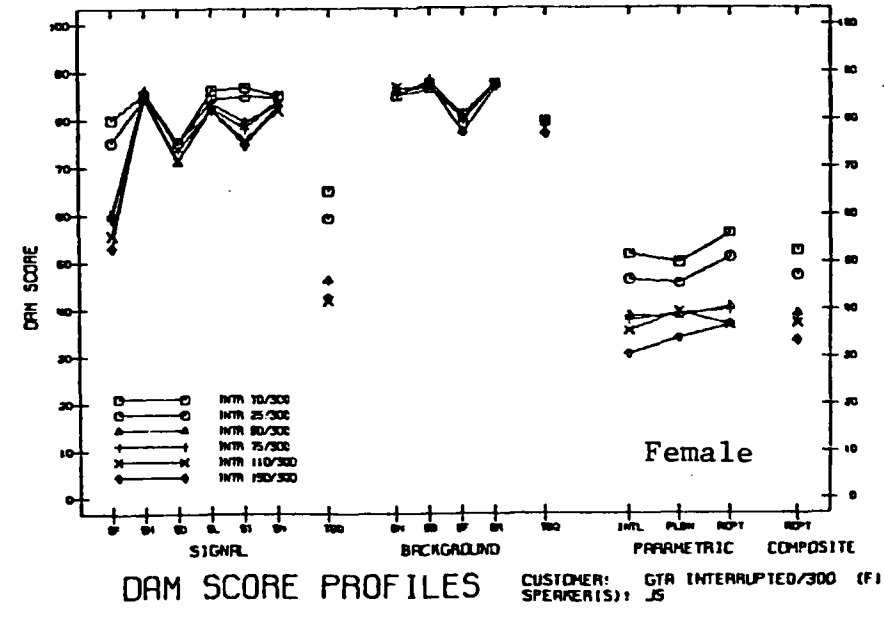
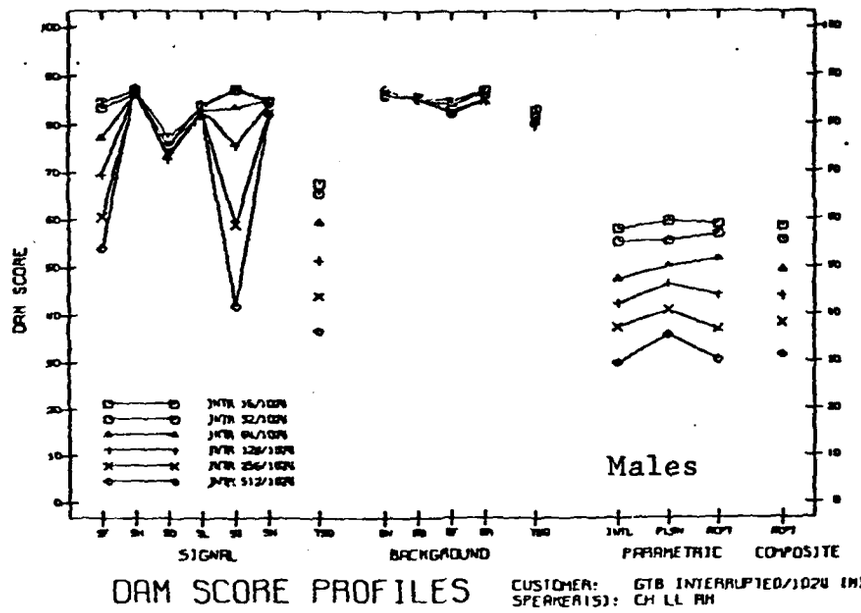


Figure 5.2.1.3-1 Effects of rapid periodic interruption on DAM scores for male and female speakers.



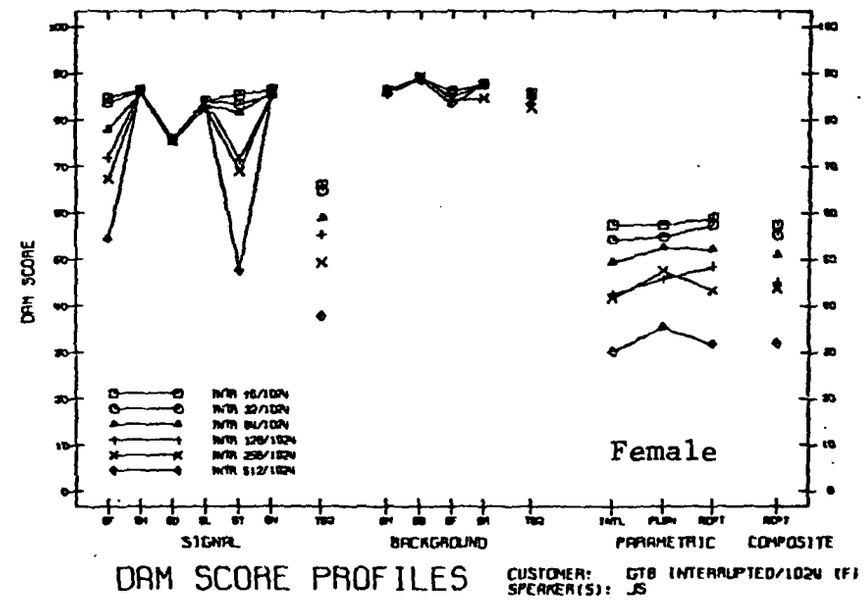
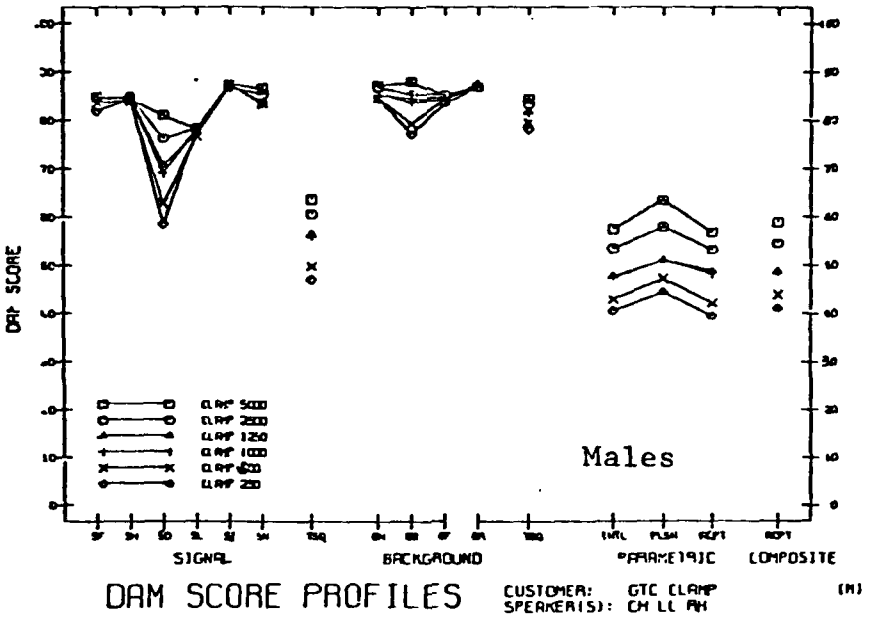


Figure 5.2.1.3-2 Effects of slower periodic interruption on DAM scores for male and female speakers.



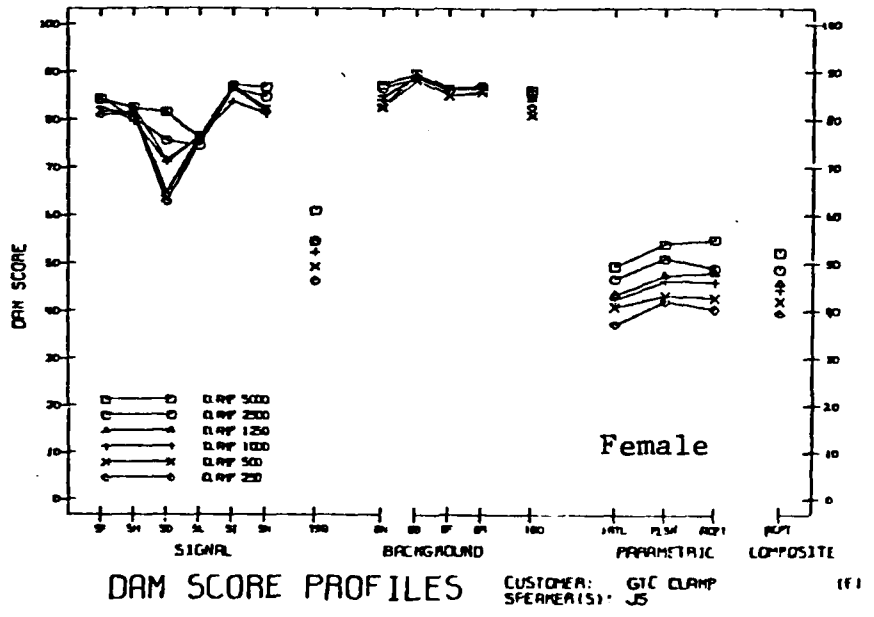


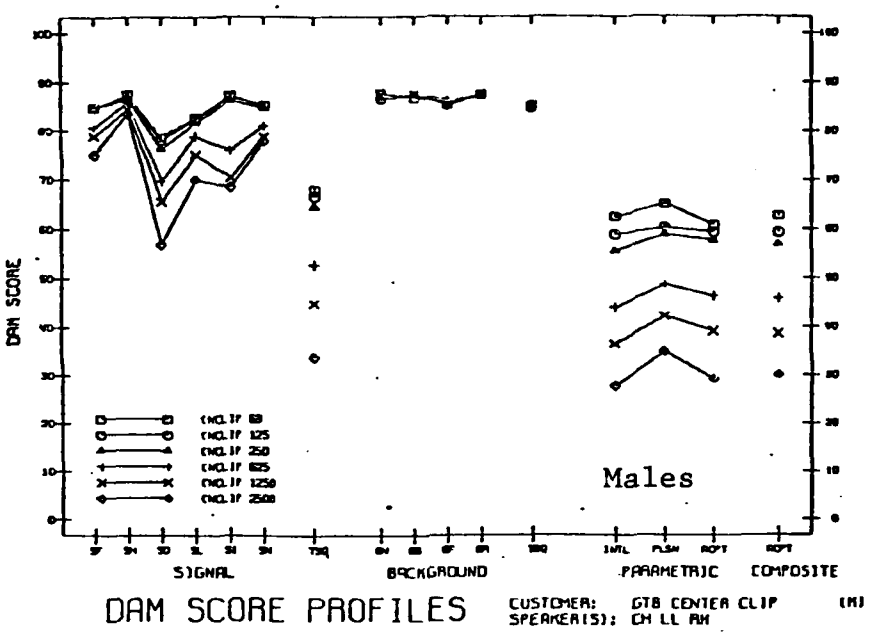
Figure 5.2.1.4 Effects of peak-clipping on DAM scores for male and female speakers.

hum, which might have been expected to increase as audio gain was increased to compensate for the effects of clipping on signal level.

5.2.1.5 Effects of center clipping

The effects of peak clipping and center clipping on speech intelligibility have long been known, but no attempt has thus far been made to quantify their effects on acceptability. Licklider [5.2] observed that peak clipping can actually enhance intelligibility under certain circumstances but that center clipping is detrimental to intelligibility under all circumstances. The reasons for this are readily found in the fact that the low energy segments of the speech signal that are removed by center clippings are generally those involving consonant sounds, which are, in turn, the major carriers of useful speech information.

Since the low energy components of speech are also those involving the upper range of the speech [illegible]s, one would predict the perceptual effects of center clipping to be considerably more complex than those of peak clipping. Fig. 5.2.1.5 shows this to be the case. From the figure it appears first that the perceptual consequences of center clipping are confined completely to perceived signal qualities. Listeners perceive virtually no background effects. Among the six signal qualities, however, the effects of center clipping are quite diverse. All but one (SH) of the signal quality scales appear highly sensitive to this form of degradation. The reasons for this diversity of effects are easily determined, moreover, once it is recalled that removing the low energy components of speech serves at once to interrupt and to low-pass the speech.



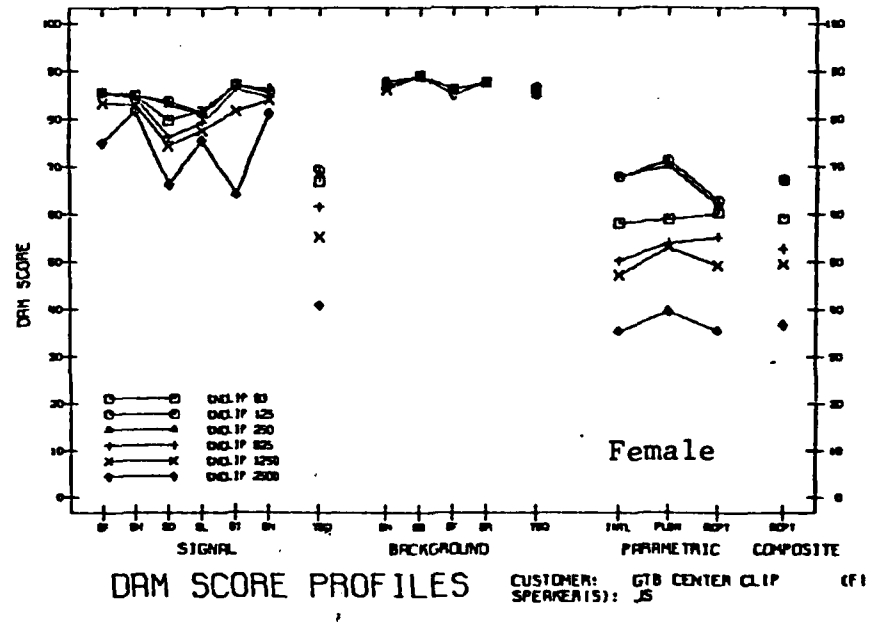


Figure 5.2.1.5 Effects of center-clipping on DAM scores for male and female speakers.

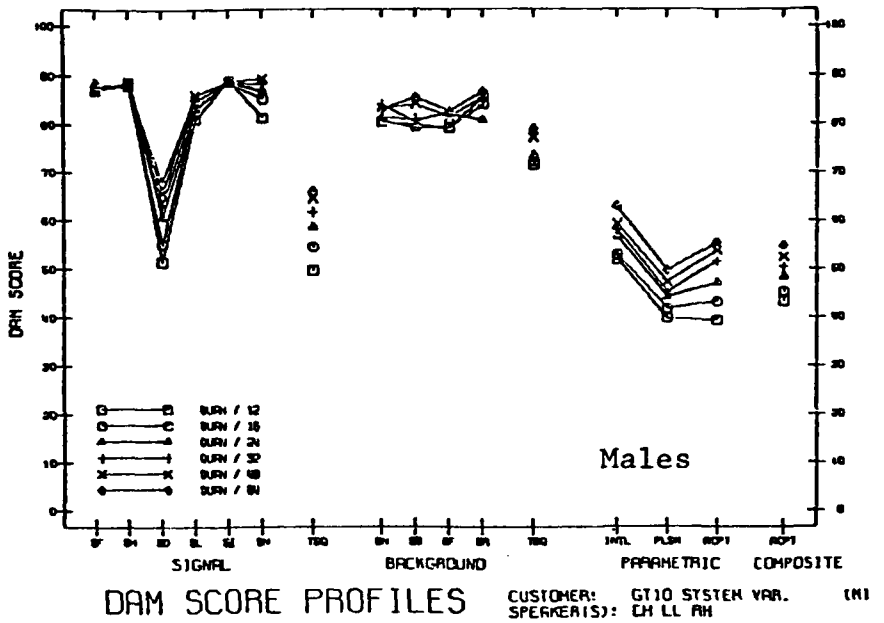
5.2.1.6 Effects of signal quantization on DAM scores

Amplitude quantization is an essential step in a number of modern speech coding techniques, though the ultimate effect in most cases is extremely fine quantization. Because the SD scale was originally designed for sensitivity to these techniques it is of some interest to know how the DAM generally and the SL scale in particular respond to relatively coarse quantization.

Figure 5.2.1.6 shows that the SL scale is in fact extremely sensitive to this form of distortion, but that several other scales are also somewhat sensitive. Perceptually, quantized speech has some of the quality of band-pass filtered speech, lowband-passed speech in particular. A moderate buzz quality is also evident. Possibly of additional interest is the difference between scores for the higher order qualities, intelligibility and pleasantness. Listeners perceive quantized speech to possess relatively high intelligibility but apparently find it unacceptable on aesthetic grounds, as evidenced by the low ratings they give it on pleasantness.

5.2.1.7 Effects of echo on DAM scores

As noted in Section 4.2.1.7 the echoic conditions treated here were somewhat unrealistic but were selected to ensure a clear subjective effect. As observed elsewhere the DAM in its present version does not make explicit provision for echo: no single rating scale pertains unequivocally to this phenomenon. From Figure 4.2.1.7, however, listeners were evidently able to find the means of distinguishing between the various delays through a combination of perceptual quality scales. It appears, moreover, that listeners experienced no uncertainty as to whether echo should be



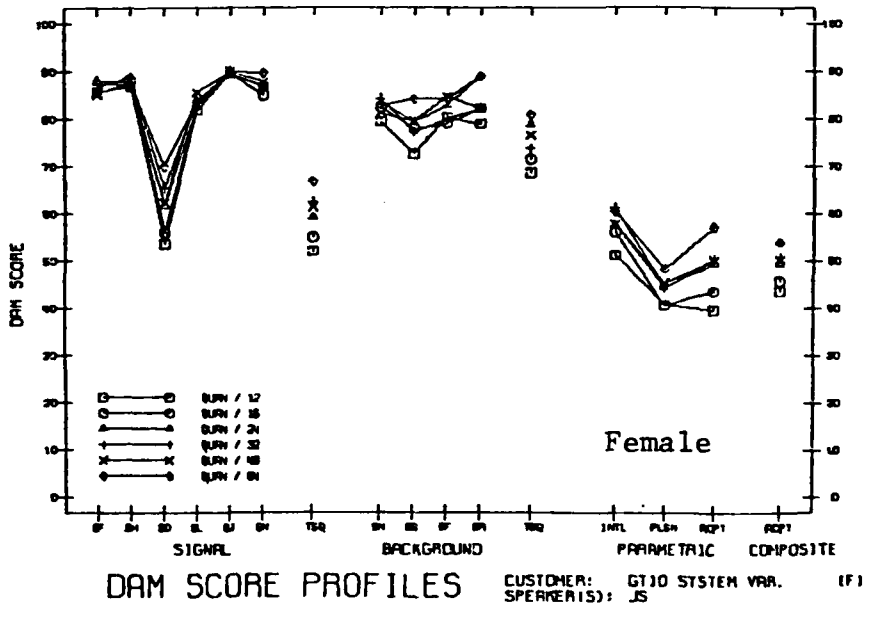


Figure 5.2.1.6 Effects of quantization on DAM scores for male and female speakers.

characterized as a signal distortion or a background disturbance. They agreed that it should be the former and indicated their perceptions primarily via the SL and SI scales, with the result that all of the higher order perceptual quality scales "tracked" in a consistent manner.

5.2.2 Effects of frequency-variant controlled distortions on DAM scores

Three classes of degradation fall in this category: additive colored noise (Section 4.2.2.2), pole distortion (Section 4.2.2.2), and banded frequency distortion (Section 4.2.2.3).

5.2.2.1 Effects of additive colored noise on DAM scores

Figure 5.2.2.1 permits a comparison of the effects of noise bands in six different frequency regions on DAM score patterns. The six bands are designated in the figure as follows:

N1 - 0- 400 Hz

N2 - 400- 800 Hz

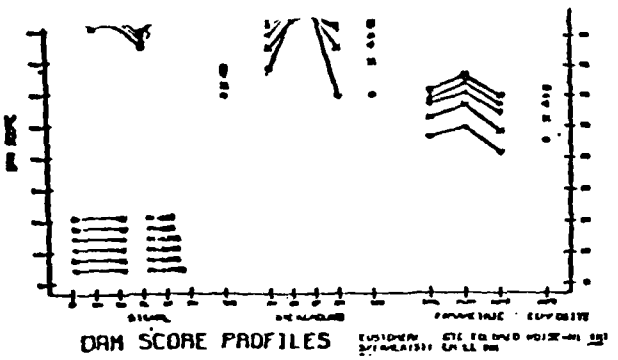
N3 - 800-1300 Hz

N4 - 1300-1900 Hz

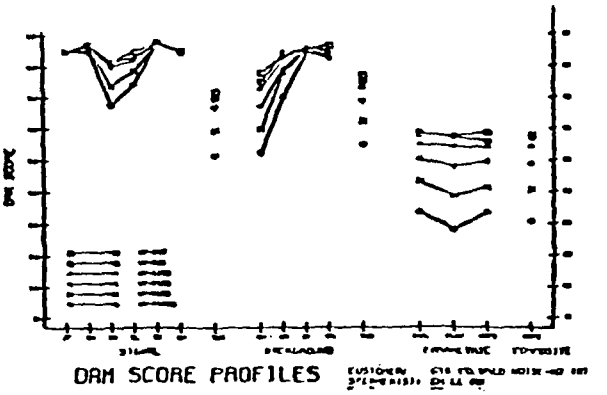
N5 - 1900-2600 Hz

N6 - 2600-3400 Hz

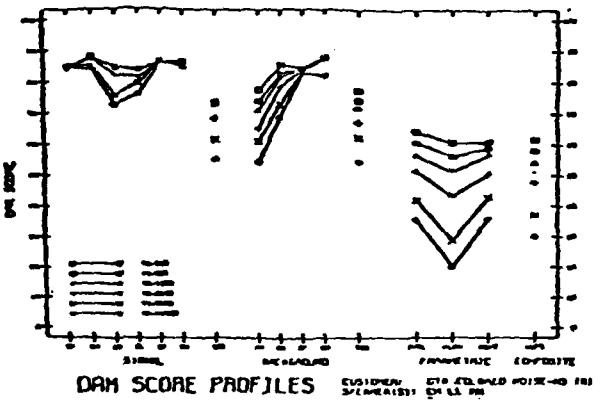
Figure 5.2.2.1M-N1 shows the pattern of DAM scores which results from speech masking by a low-frequency band of noise. Depressed scores on the BN (background noise) and BR (background rumble) scales conform with intuitive expectations, and otherwise serve to provide additional validation of these two scales. Not immediately clear is the reason for somewhat depressed scores on several perceived signal quality scales and on the scale for total signal quality.



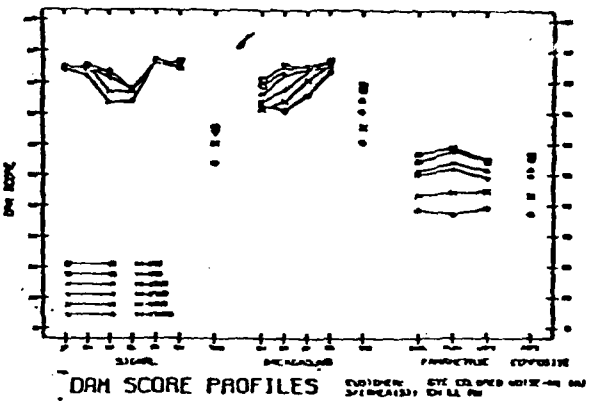
N1: 0-400 Hz noise



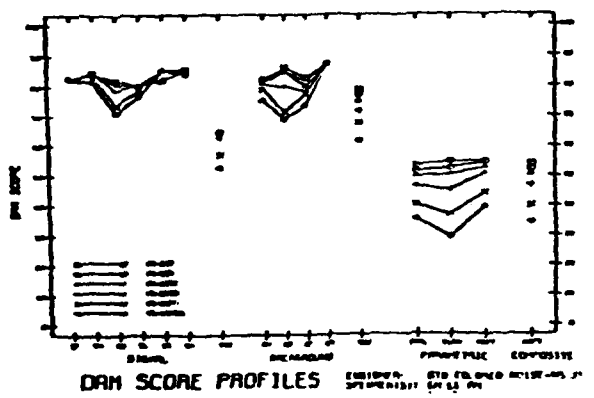
N2: 400-800 Hz noise



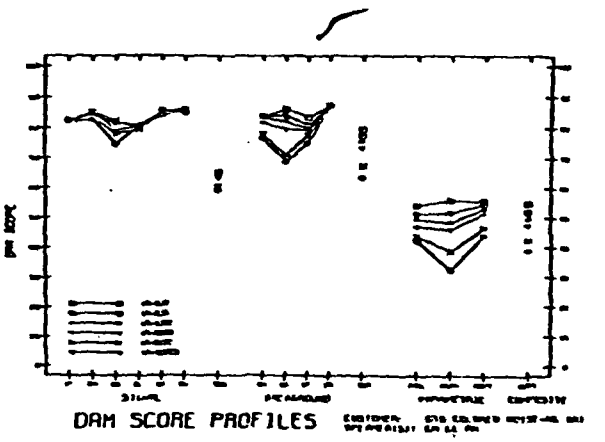
N3: 800-1300 Hz noise



N4: 1300-1900 Hz noise



N5: 1900-2600 Hz noise



N6: 2600-3400 Hz noise

Figure 5.2.2.1M Effects of narrow-band noise on DAM scores for male speakers.

As shown in Figure 5.2.2.1M-N2, the quality, background rumble, decreases significantly as the noise band is raised above the 0-400 Hz region.

As the frequency region of the noise band is increased beyond the 800-1300 Hz frequency region, the perceptual consequences of the noise undergo several qualitative changes. The noise is perceived to have less "rushing-roaring" quality (BN) but more of a "buzzing-humming" quality as reflected in scores on the BB scale. At the highest noise levels listeners also tend to perceive an increasing raspy (SD scale) quality which is most typical of amplitude-distorted speech.

5.2.2.2 Effects of Pole Distortion on DAM scores

Two types of pole distortion, as described in Section 4.2.2.2, are examined in this section. The first of these involves distortion of pole frequencies within a given frequency band, the second, involves "radial distortion" and, hence, band-width distortion.

5.2.2.2.1 Effects of pole frequency distortion

Figure 5.2.2.2-1 shows the effects of pole distortion in each of six frequency bands:

200- 400 Hz

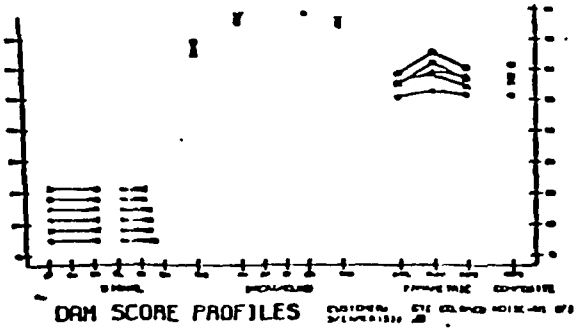
400- 800 Hz

800-1300 Hz

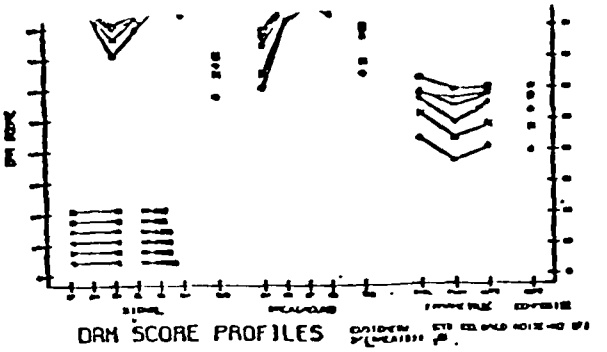
1300-1900 Hz

1900-2600 Hz

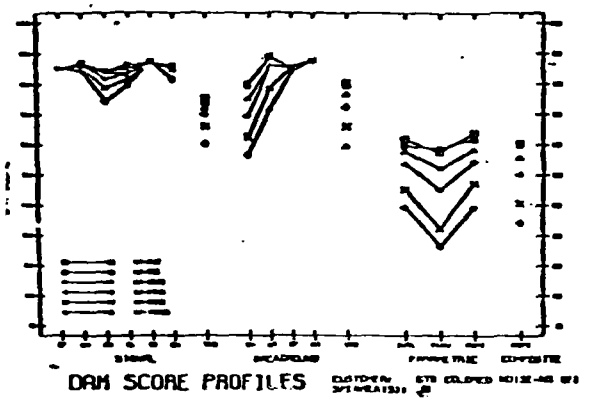
2600-3400 Hz



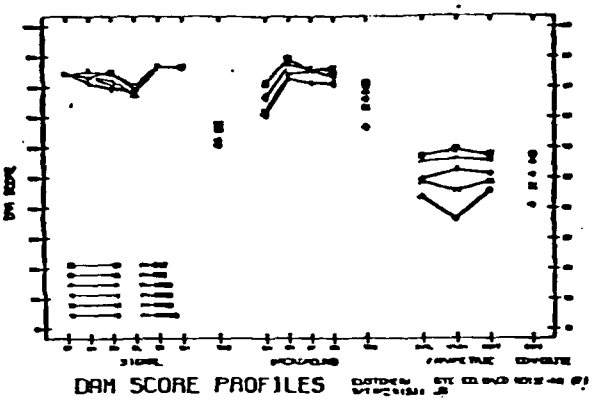
N1: 0-400 Hz noise



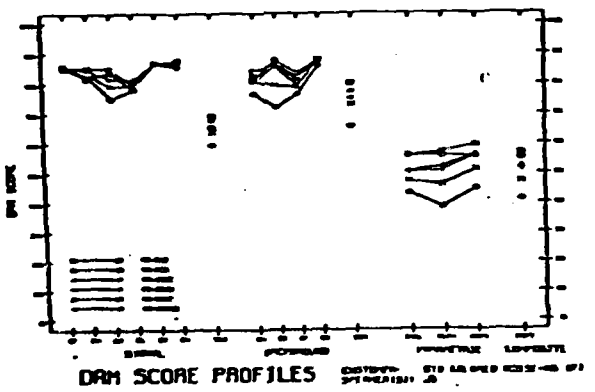
N2: 400-800 Hz noise



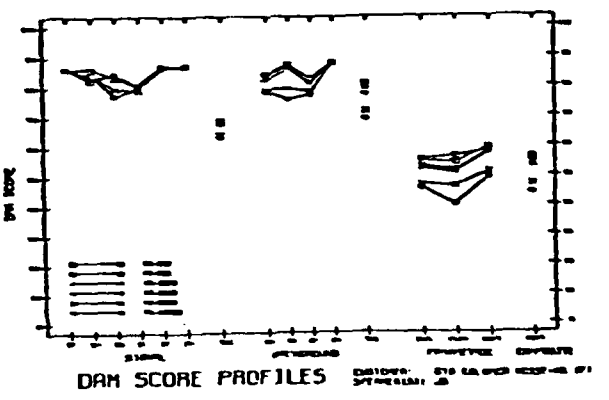
N3: 800-1300 Hz noise



N4: 1300-1900 Hz noise

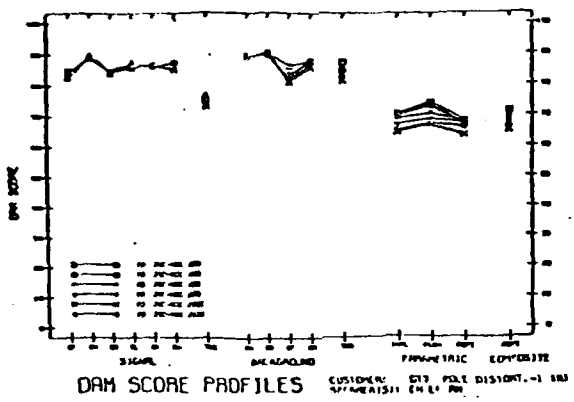


N5: 1900-2600 Hz noise

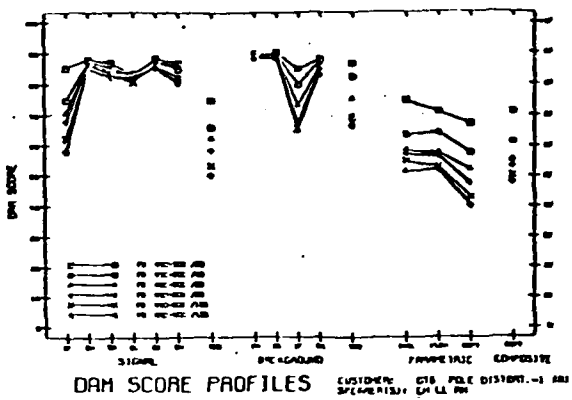


N6: 2600-3400 Hz noise

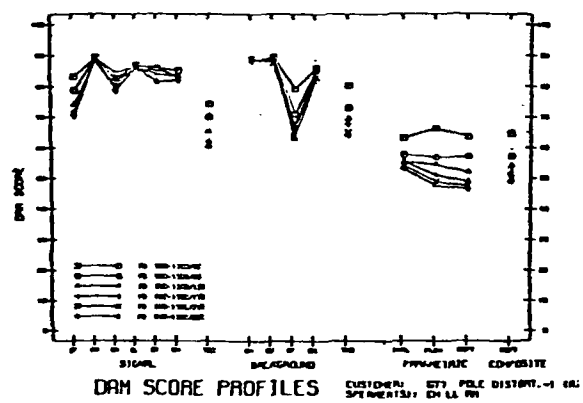
Figure 5.2.2.1 F Effects of narrow-band noise on DAM scores for a female speaker.



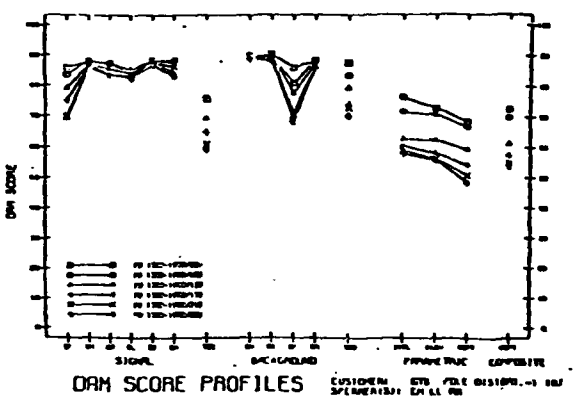
Band 1: 200-400 Hz



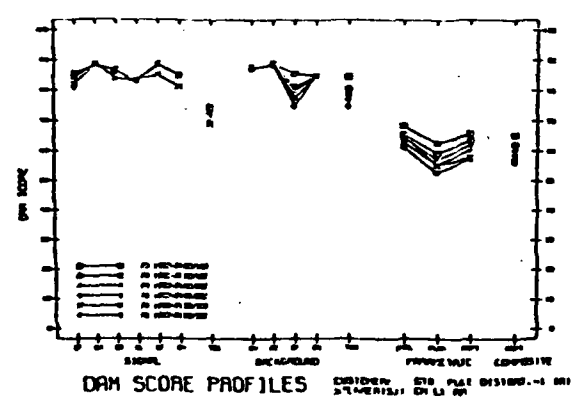
Band 2: 400-800 Hz



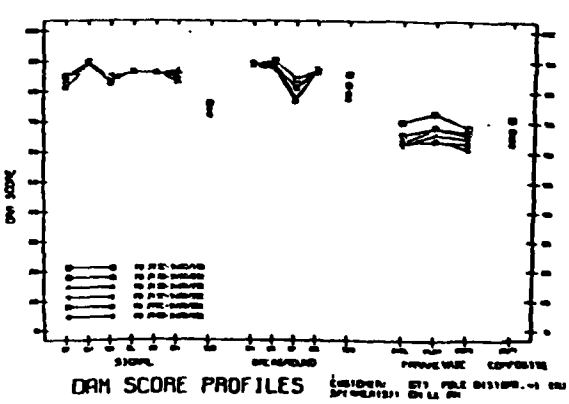
Band 3: 800-1300 Hz



Band 4: 1300-1900 Hz

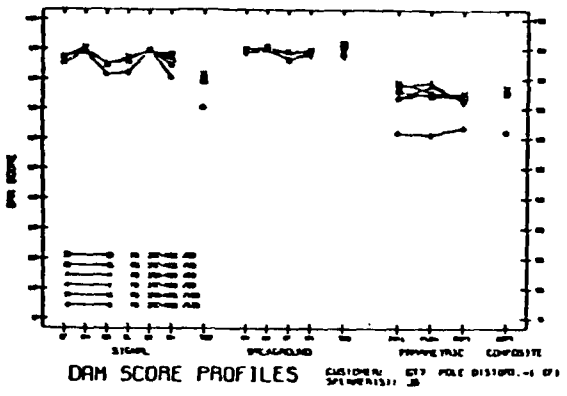


Band 5: 1900-2600 Hz

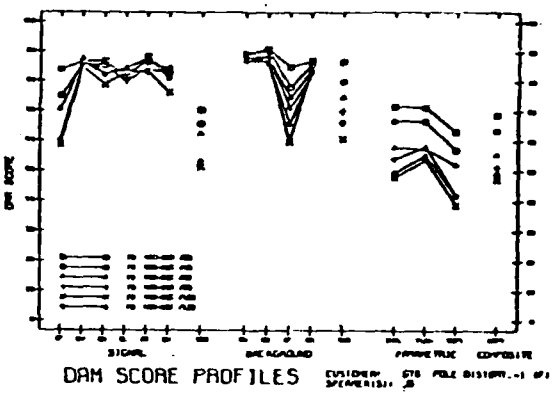


Band 5: 2600-3400 Hz

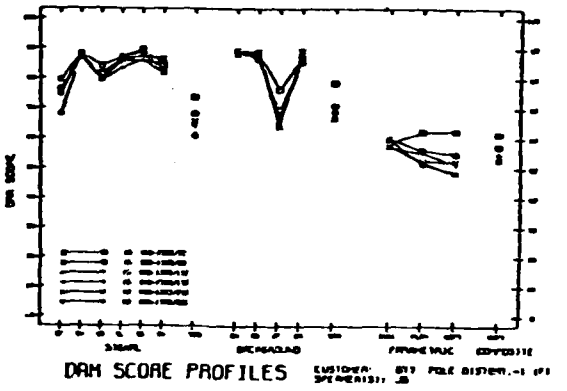
Figure 5.2.2.2-1M Effects of pole-frequency distortion on DAM scores for male speakers.



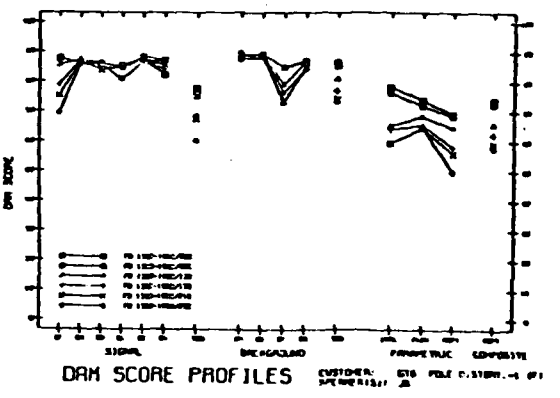
Band 1: 200-400 Hz



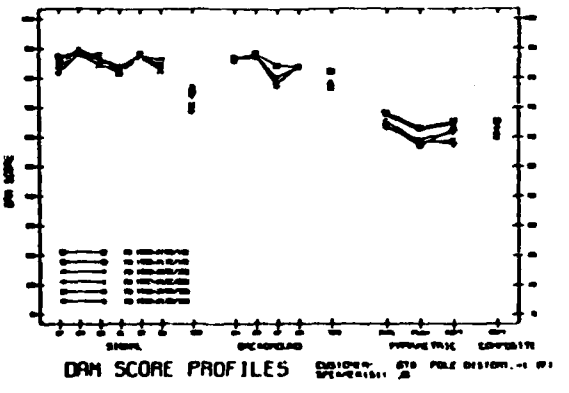
Band 2: 400-800 Hz



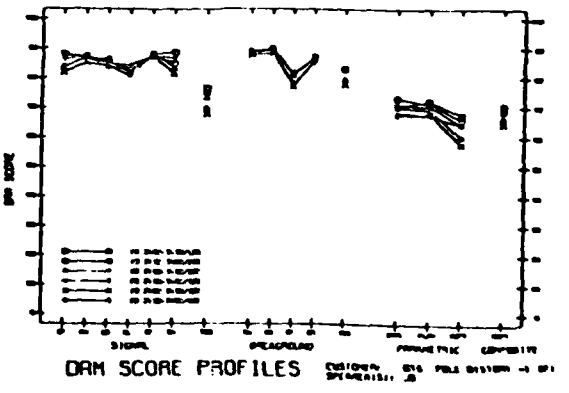
Band 3: 800-1300 Hz



Band 4: 1300-1900 Hz



Band 5: 1900-2600 Hz



Band 6: 2600-3400 Hz

Figure 5.2.2.2-1F Effects of pole-frequency distortion on DAM scores for female speaker

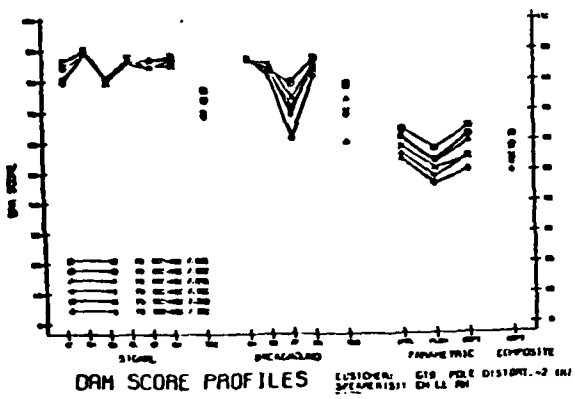
The parameter within each sub-figure is range of frequency distortion (rms). Values of this parameter are as follows:

BAND					
0-400	400-800	800-1300	1300-1900	1900-2600	2600-3400
20	20	50	50	100	150
40	40	90	90	150	200
60	60	130	130	200	250
80	80	170	170	250	300
100	100	210	210	300	350
120	120	250	250	350	400

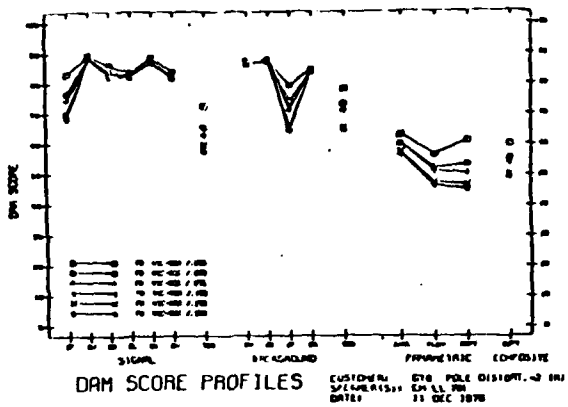
From Figure 5.2.2.2-1, it appears that the subjective effects of pole frequency distortion are expressed primarily via the SF (signal flutter) and BF (background flutter) scales. The remaining perceptual quality scales are virtually unaffected by this form of degradation. It appears, farther, that the perceptual consequences of pole distortion are generally neglible in the upper end lower extremes of the 3.4K Hz band involved here.

5.2.2.2-2 Effects of radial pole distortion

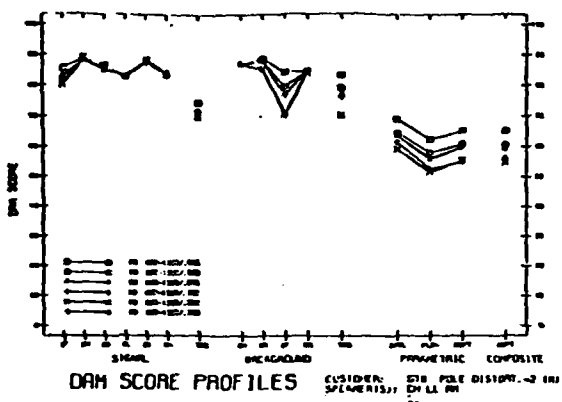
Figure 5.2.2.2-2 shows the effects of radial pole distortion. In this case the frequency bands involved were as indicated above except for the lowest band which was 0-400 Hz instead of 200-400 Hz. The parameter in all sub-figures is relative "radius jitter." The values being:



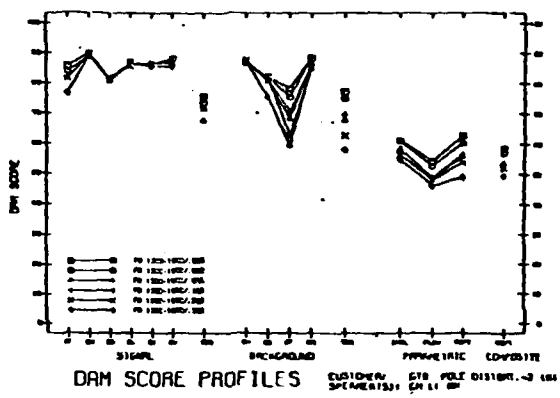
Band 1: 0-400 Hz



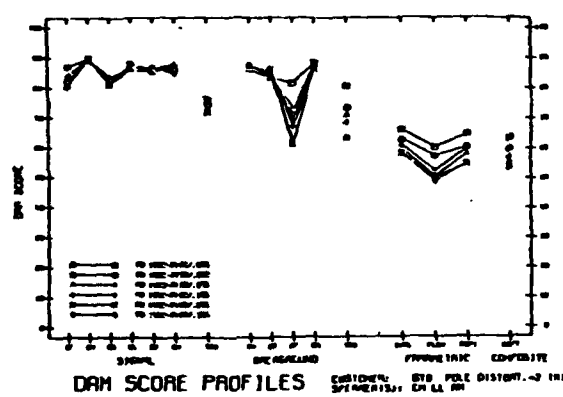
Band 2: 400-800 Hz



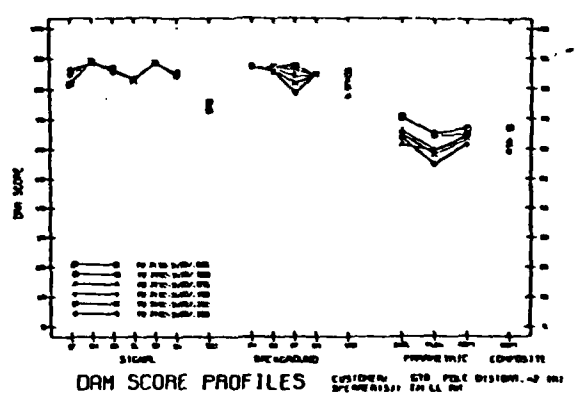
Band 3: 800-1300 Hz



Band 4: 1300-1900 Hz

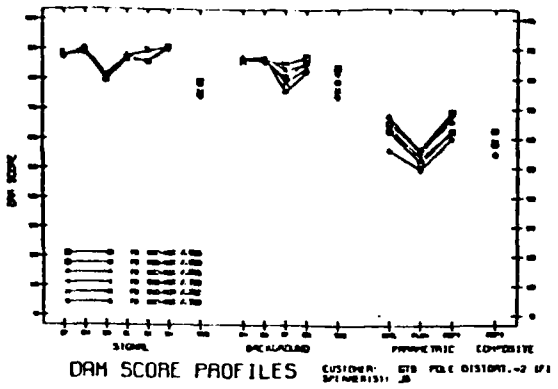


Band 5: 1900-2600 Hz

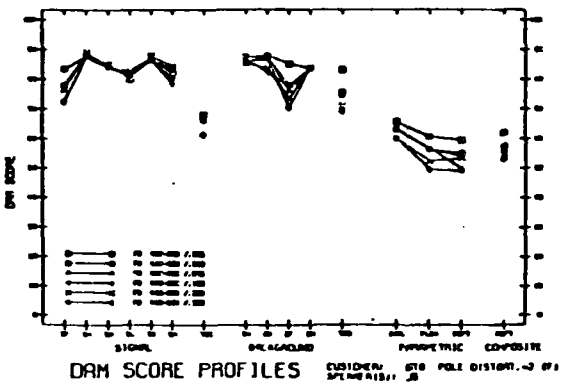


Band 6: 2600-3400 Hz

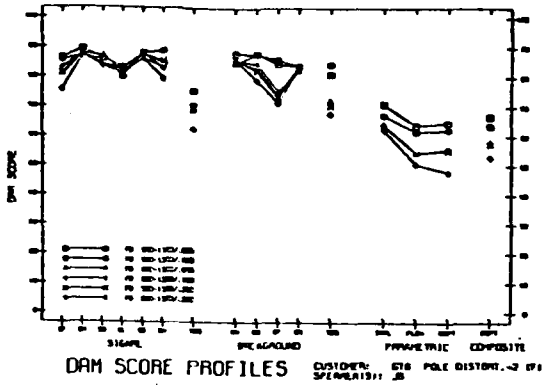
Figure 5.2.2.2-2M Effects of radial pole distortion on DAM scores for male speakers.



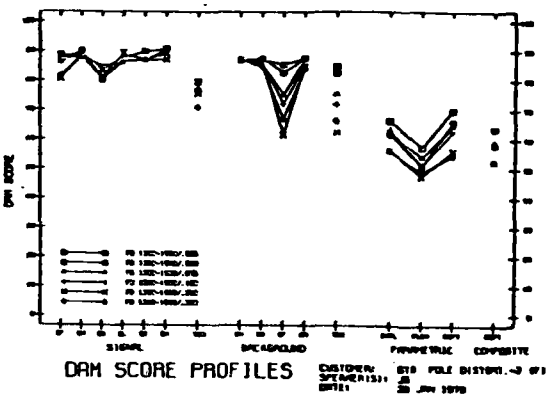
Band 1: 0-400 Hz



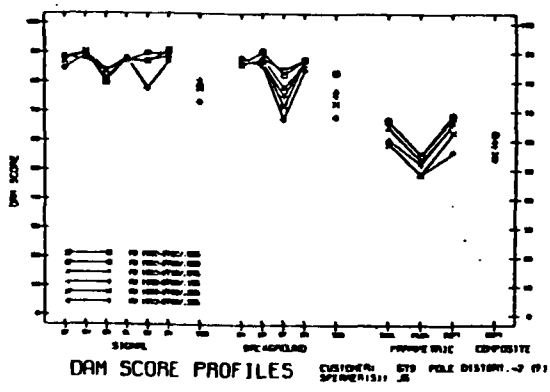
Band 2: 400-800 Hz



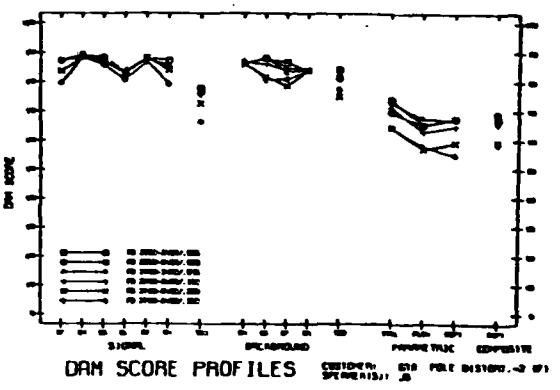
Band 3: 800-1300 Hz



Band 4: 1300-1900 Hz



Band 5: 1900-2600 Hz



Band: 2600-3400 Hz

Figure 5.2.2.2-2F Effects of radial pole distortion on DAM scores for female speaker.

.025

.050

.075

.100

.200

.300

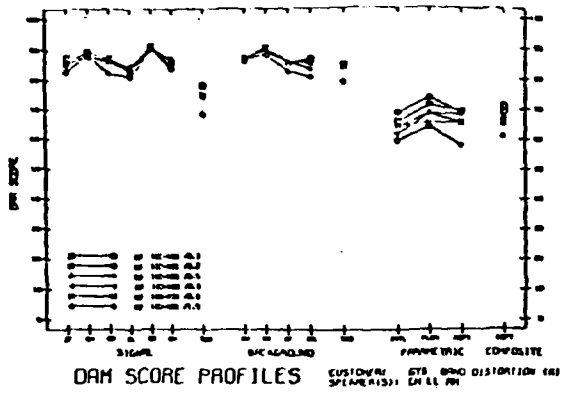
in all cases. From the figure it is evident pole-bandwidth distortion has qualitatively different perceptual consequences than pole-frequency distortion. The perceptual effects of radial or bandwidth distortion are confined primarily to the background and are quite pronounced in all but the highest frequency band.

5.2.2.3 Banded frequency distortion

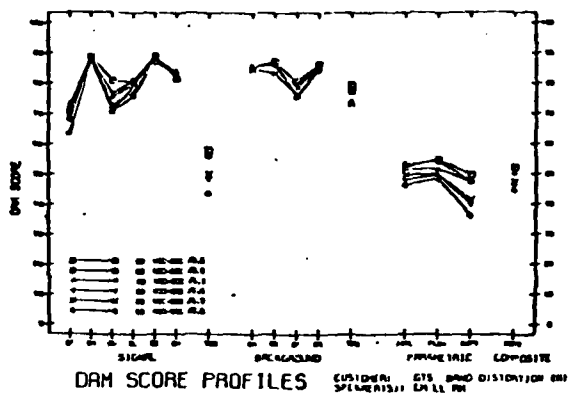
Banded frequency distortion is of interest in relation to transform coding techniques where noise may be a factor at the power spectral level. In the present case six levels of noise were produced in each of six spectral bands (See Section 4.2.2.3).

From Figure 5.2.2.3 it appears that banded frequency distortion in the range treated here has relatively minor subjective consequences. In all but the lowest frequency band involved, perceived background flutter is the most pronounced effect.

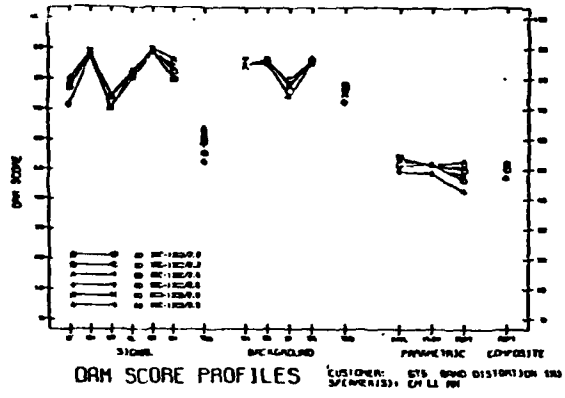
Some amount of signal flutter (SF scale) and raspiness (SD scale) is evident in the cases of the 400-800 Hz and 800-1300 Hz bands, but these qualities are negligible in the remaining bands.



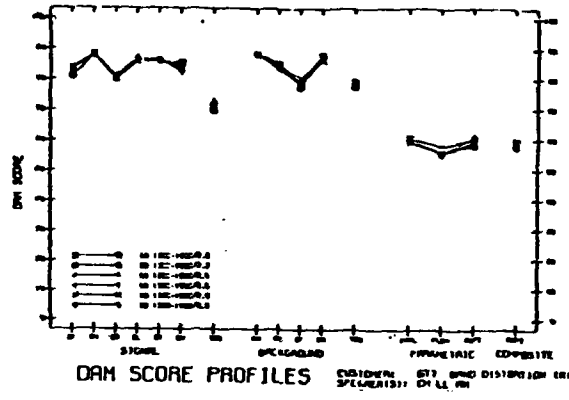
Band 1: 100-400 Hz



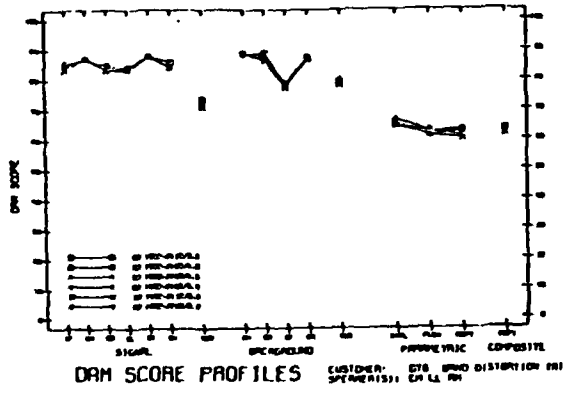
Band 2: 400-800 Hz



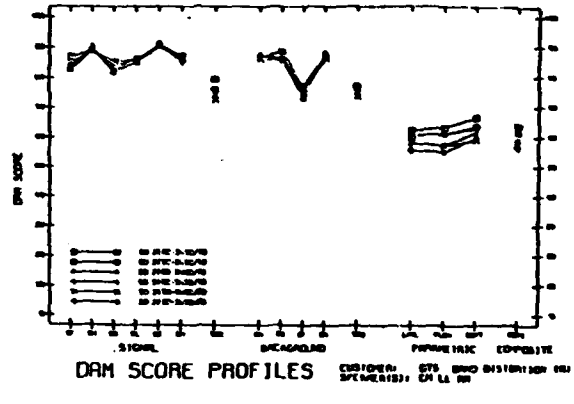
Band 3: 800-1300 Hz



Band 4: 1300-1900 Hz

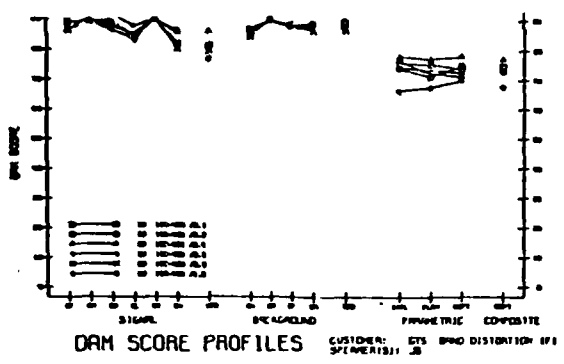


Band 5: 1900-2600 Hz

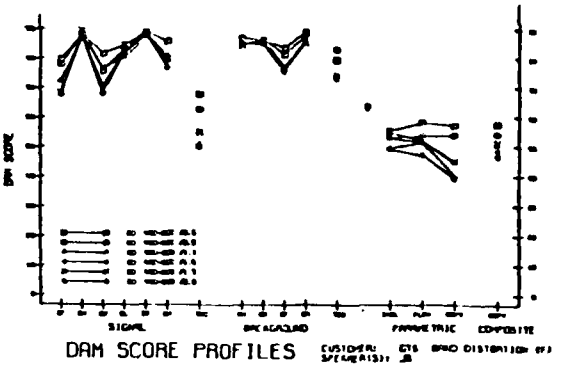


Band 6: 2600-3400 Hz

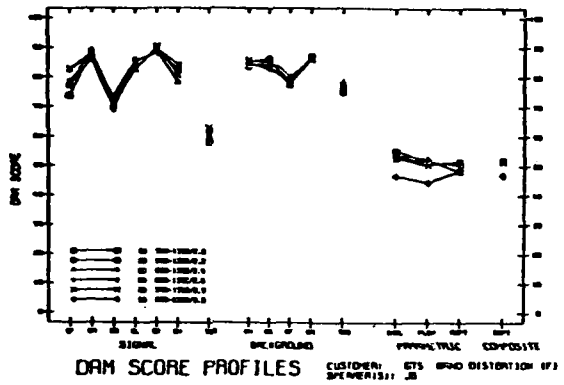
Figure 5.2.2.3M Effects of banded frequency distortion on DAM scores for male speakers.



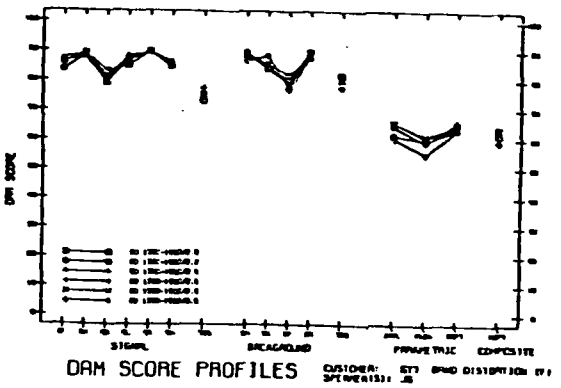
Band 1: 100-400 Hz



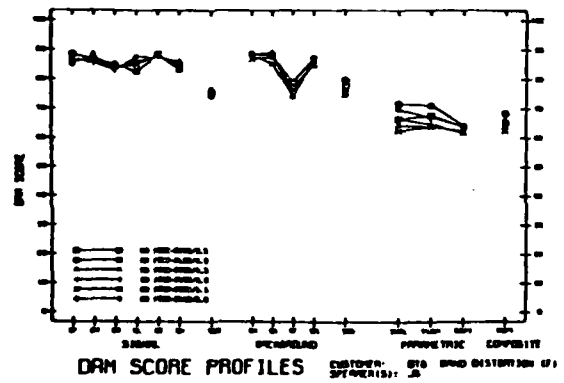
Band 2: 400-800 Hz



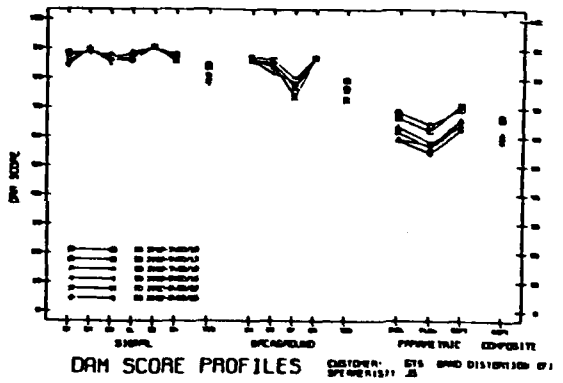
Band 3: 800-1300 Hz



Band 4: 1300-1900 Hz



Band 5: 1900-2600 Hz



Band 6: 2600-3400 Hz

Figure 5.2.2-3F Effects of banded frequency distortion on DAM scores for a female speaker.

REFERENCES

5.1 W. D. Voiers, "Diagnostic Acceptability Measure for Speech Communications System," Conference Record, ICASSP, Hartford, CN, 1977.

5.2 Licklider, "Effects of Amplitude Distortion Upon the Intelligibility of Speech," JASA, Vol. 18, No. 2, 1946.

CHAPTER 6. THE EXPERIMENTAL RESULTS

6.1 Introduction

This chapter gives both a detailed description of the experiments performed as part of this study and a complete analyses of the experimental results. In all cases, the experiments performed were based on correlation analyses, and the figure-of-merit used for each objective quality measure studied was either the estimated correlation coefficient, $\hat{\rho}$, or the estimate standard deviation of error when the objective measure is used to estimate the subjective measure, $\hat{\sigma}_e$ (see Chapter 1).

This chapter is divided into five additional sections. The first describes the standard analysis techniques used in this study. The second describes the results of the spectral distance measure studies. The third describes the results from the parametric measure studies. The fourth describes the results from the frequency variant measure studies. The fifth describes the results from the composite measures.

6.2 Analysis Procedures

Every correlation experiment performed as part of this study resulted in an estimated correlation coefficient, $\hat{\rho}$, and an associated estimated standard deviation of error, $\hat{\sigma}_e$. To describe each experiment exactly, one must therefore know four things: exactly what objective measure(s) was used; exactly what analysis method was used; exactly what subjective parameter was used; and exactly what set of distortions was used in the correlation analyses. The first three items will be discussed in

this section. The objective measures will be discussed in the following sections.

6.2.1 The Estimation Procedures

The three estimation procedures used in this study were linear regression, non-linear regression, and linear multi-regression. In linear regression, the subjective result is estimated from the objective result by

$$\hat{S}(s,d) = \beta(1)\ O(s,d) + \beta(0) \qquad 6.2.1\text{-}1$$

where $\hat{S}(s,d)$ is the estimate of the subjective result for speaker s and distortion d, $O(s,d)$ is the objective measure for speaker s and distortion d, and $\beta(1)$ and $\beta(0)$ are constants. The solution which gives a minimum squared error between $S(s,d)$ and $\hat{S}(s,d)$ is given by

$$\beta(1) = \frac{\hat{\rho}\hat{\sigma}_s}{\hat{\sigma}_0} \qquad 6.2.1\text{-}2$$

and

$$\beta(0) = \bar{S} - \bar{O}\,\frac{\hat{\rho}\hat{\sigma}_s}{\hat{\sigma}_0} \qquad 6.2.1\text{-}3$$

where $\hat{\sigma}_s$ is the estimated standard deviation over the subjective data, $\hat{\sigma}_0$ is the estimated standard deviation over the objective data, $\hat{\rho}$ is the estimated correlation coefficient, $\bar{S}$ is the average subjective result, and $\bar{O}$ is the average objective result.

In non-linear regression analysis, the subjective estimate is given by

$$\hat{S}(s,d) = \sum_{k=0}^{K} \beta(k)\ O^k(s,d) \qquad 6.2.1\text{-}4$$

where K is the order of the regression. Note that for K=1, this equation becomes the linear regression equation. To find $\beta(k)$, the subjective error, given by

$$E^2(s,d) = (S(s,d)-\hat{S}(s,d))^2 \qquad 6.2.1\text{-}5$$

is minimized with respect to $\beta(k)$. This leads to a set of linear equations of the form

$$\underline{\underline{E}}\ \beta = \underline{Z} \qquad 6.2.1\text{-}6$$

where

$$\underline{\beta}^T = [\beta(0), \beta(1), \ldots, \beta(N)] \qquad 6.2.1\text{-}7$$

$$\underline{Z}^T = [\sum_{s,d} s(s,d), \sum_{s,d} O(s,d)S(s,d), \ldots, \sum_{s,d} O^k(s,d)S(s,d)] \qquad 6.2.1\text{-}8$$

and

$$E(k,\ell) = \sum_{s,d} O^{k+\ell}(s,d) \qquad 6.2.1\text{-}9$$

where $E(k,\ell)$ is the k and ℓ entry to the matrix E. Once E is inverted, $\hat{\rho}$ is obtained from

$$\hat{\sigma}_{\hat{s}}^2 = \frac{1}{N-1}[\underline{\beta}^T\ \underline{\underline{E}}\ \underline{\beta} - N \sum_{k=0}^{K} \beta(k)\overline{O}^k] \qquad 6.2.1\text{-}10$$

$$\hat{\rho} = \frac{\left[\frac{1}{N-1}\left(\underline{\beta}^T \underline{Z} - N \bar{S} \sum_{k=0}^{K} \beta(k)\bar{O}^k\right)\right]^{1/2}}{\hat{\sigma}_s \hat{\sigma}_s} \qquad 6.2.1\text{-}11$$

where N is the total number of points in the sample.

Linear multiregression is in many ways similar to non-linear regression. In this procedure, it is desired to estimate the subjective results from several (K) different objective measures by

$$\hat{S}(s,d) = \sum_{k=0}^{K} \beta(k) O(s,d,k) \qquad 6.2.1\text{-}12$$

where the extra index "k" has been added to the objective measure to differentiate the different measures. To find $\beta(k)$, the squared subjective error

$$\sum_{s,d} e^2(s,d) = \sum_{s,d} (S(s,d)-\hat{S}(s,d))^2 \qquad 6.2.1\text{-}13$$

is minimized, giving

$$\underline{\underline{e}}\,\underline{\beta} = \underline{Z} \qquad 6.2.1\text{-}14$$

where

$\underline{\beta}^T$ is given, as before, by equation 6.2.1-7, $\underline{Z}^T$ is given by

$$\underline{Z}^T = \left[\sum_{s,d} S(s,d), \sum_{s,d} S(s,d)O(s,d,1), \ldots, \sum S(s,d)O(s,d,k)\right] \qquad 6.2.1\text{-}15$$

and

$$e(k,\ell) = \sum_{s,d} O(s,d,k)O(s,d,\ell) \qquad 6.2.1\text{-}16$$

After $\underline{\beta}$ is computed by inverting $\underline{\underline{e}}$, $\hat{\rho}$ may be computed from

$$\hat{\sigma}_{\hat{s}}^2 = \frac{1}{N-1}\left[\underline{\beta}^T \underline{\underline{e}} \underline{\beta} - N \sum_{k=0}^{K} \beta(k)\overline{O(s,d,k)}\right] \qquad 6.2.1\text{-}17$$

and

$$\hat{\rho} = \frac{\frac{1}{N-1}\left[\beta^T Z - N \bar{S} \sum_{k=0}^{K} \beta(k)\overline{O(s,d,k)}\right]}{\hat{\sigma}_{\hat{s}}\ \hat{\sigma}_{\hat{s}}} \qquad 6.2.1\text{-}18$$

where $O(s,d,0) = 1$.

6.2.2 The Distorted Data Sets

In Chapter 4, a detailed description of the distorted data base was given. This data base contained coding distortions, wideband controlled distortions, and frequency variant controlled distortions. There are several points which should be made about this data base. First, it was heavily loaded with frequency variant distortions because it was felt that considerable improvement in objective quality measures might be achieved by better understanding the frequency variant perceptual effects. Hence, measures tested over the set of all distortions, called ALL, is of considerable interest, and represents a lower limit on the performance of any measure.

However, an ensemble of distortions which contains as many frequency variant distortions as this data base does not represent a good estimate of a true coding environment. Hence, a second major distortion set was identified, called WBC (wide band distortions) which, in the opinion of the researchers, gives a better estimate of the true behavior of the measures in a true coding environment. A description of the distortion set WDB is given in Table 6.2.2-1.

In addition to WDB, a total of seven additional data subsets were identified and used. These were WFC (waveform coders), CODE (coding

Coding Distortions	# of Cases	WBD	WFC	CODE	CON	WBN	NBN	BD	PD	ND
ADPCM	6	X	X	X						
APCM	6	X	X	X						
CVSP	6	X	X	X						
ADM	6	X	X	X						
APC	6	X	X	X						
LPC	6	X		X						
VEV	12	X		X						
ATC	6	X	X	X						
Controlled Distortion										
Additive Noise	6	X	X		X	X				X
Low pass filter	6	X			X					
High pass filter	6	X			X					
Band pass filter	6	X			X					
Interruption	12	X			X					
Clipping	6	X			X					
Center clipping	6	X			X					
Quantization	6	X	X	X	X					
Echo	6				X					
Frequency Variant										
Additive colored noise	36				X		X			X
Banded pole distortion	78				X				X	
Banded frequency distortion	36				X			X		

Table 6.2.2-1. SUBCLASSES OF DISTORTIONS USED AS PART OF THIS RESEARCH

distortions), CON (controlled distortion), WBN (wide band noise), NBN (narrow band noise), BD (band distortion), and PD (pole distortions). The contents of these various sets are also shown in Table 6.2.2-2.

6.2.3 The Subjective Data Set

In all, the subjective data base contains 20 subjective results per distortion. Although 18 of these were used in the total data analysis of this study, the emphasis in on the results on only a few. This includes CA (composite acceptability), TBQ (total background quality), and TSQ (total system quality) for the isometric measures, and all the parametric results for the parametric measures. Of these, CA was considered most important, and most major isometric results are based on this measure.

6.2.4 Non-parametric Rank Statistics

An important part of this study was the comparison of different analysis methods and parameterizations for their ability to better predict subjective results. Based on our figures-of-merit, correlation coefficients and standard deviation of error, it is easy to rank these methods with respect to one another. The problem is that the specific statistical environment for our tests, namely correlation coefficient estimates with non-zero centered correlation coefficients across correlated sample sets, has not been widely treated in the literature.

In order to get some statistical handle on this problem, non-parametric pairwise rank statistics were used. In this approach, treatments are always treated in pairs, so that the question being asked is always if one treatment is better than the other. The data base is then scanned to find all cases where two measures differ only in that one of the measures has received treatment 1 and the other has received treatment 2.

The null hypothesis is that the treatments make no difference. If this were true, then each of the treatments would be ranked first in the pairs in about one-half of the cases. Let there be N such cases, and let the rank of the first treatment (either 1 or 2) be given by RK(1,n), $1 \leq n \leq N$. Then the rank statistic which is formed, called RS, is given by

$$RS = \frac{1}{N} \sum_{n=1}^{N} RK(1,n) \qquad 6.2.4\text{-}1$$

This statistic varies between 1 and 2. If it is equal to 1, then the first treatment was always ranked first. If it is equal to 2, then the first treatment is always ranked second.

RS can only take on a finite set of values, namely $\frac{N}{N}$, $\frac{N+1}{N}$, ..., $\frac{2N-1}{N}$, $\frac{2N}{N}$. The probability is that RS takes on a value $\frac{N+\alpha}{N}$ is given by

$$\text{prob}\left(\frac{N+\alpha}{N}\right) = \frac{\frac{N!}{\alpha!\,(N-\alpha)!}}{2^N} \qquad 6.2.4\text{-}2$$

Hence, the probability that RS takes on a value of $\left(\frac{N+\alpha}{N}\right)$ or less is given by

$$\text{prob}\left(RK \leq \frac{N+\alpha}{N}\right) = \frac{1}{2^N} \sum_{a=0}^{\alpha} \frac{N!}{a!\,(N-a)!} \qquad 6.2.4\text{-}3$$

From this relationship, it is always easy to compute the significance of a ranking in the usual sense.

For multiple values of the same parameter (i.e., multiple treatments of the same type), all possible pairwise rankings were done. An example of the results of such an analysis for four parameter values is given in Table 6.2.4-1. Above the diagonal in the matrix is placed the

	PARAMETERS			
	1	2	3	4
1	----	RS12	RS13	RS14
2	SL12(N12)	----	RS23	RS24
3	SL13(N13)	SL23(N23)	----	RS34
4	SL14(N14)	SL24(N24)	SL34(N34)	----

RSXY = Rank statistic between parameters X & Y (equ. 6.2.4-1)

SLXY = Significance limit (in the probability domain) for the X-Y rank statistic

NXY = Number of samples available for computing RSXY

Table 6.2.4-1. EXAMPLE LAYOUT FOR THE RESULTS OF A FOUR PARAMETER PAIRED RANKING TEST

pairwise values for RS. Below the diagonal is placed the one-sided probability limit. For significance at the .01 level, this number must be below .01, and for significance at the .05 level, it must be below .05.

The pairwise ranking test described here is a relatively weak statistical test. It has been adopted because it does give some statistical insight into the significance of the test results, and because many of the results reported here are very strong.

6.3 The Spectral Distance Measure Results

A total of 192 variations of the spectral distance measures described in Chapter 3 were included as part of this study. Any of these spectral distance measures can be described by four conditions. First, the spectral distance measure may be between linear spectra, log spectra, or a spectrum taken to the δ power. If the latter case is used, the value of δ must be specified. Second, between frames, the measures are weighted by the energy of the original signal taken to the α power. If α=0, then there is no energy weighting. Third, the measures always involve an L_p norm, and the value of p is important. Fourth, within frames, the distance measure may be spectrally weighted by $V(n,s,d,\theta_\ell)^\gamma$. If γ=0, there is no spectral weighting. In these terms, Table 6.3-1 summarizes the 192 spectral distance measures studied here.

The total analysis performed on the 192 spectral distance measures was linear, 3rd order nonlinear and 6th order nonlinear regression. These analyses were performed across all nine of the distortion subsets (ALL, WBD, WFC, CODE, CON, WBN, NBN, BD, PD) for nine subjective parameters (CA, TBQ, TSQ, P, A, I, PP, PA, PI). In all, there were therefore 192 x 3 x 9 x 9 = 46,656 analyses. Obviously, it is unreasonable to even print this

SUMMARY OF SPECTRAL DISTANCE MEASURES

Linear Spectral Distance Measures

Energy Weighting (α)	0	.5	1	2				
Lp Norm (P)	1	2	4	8				
Spectral Weighting (γ)		0	1	2				
Total cases = 48								

Log Spectral Distance Measures

Energy Weighting (α)	0	.5	1	2				
Lp Norm (P)	1	2	4	8	10	12	14	16
Spectral Weighting (γ)		0	1	2				
Total cases = 64								

Spectral Distance Measures

Energy Weighting (α)	0							
Lp Norm (P)	1	2	4	8	10	12	14	16
Spectral Weighting (γ)		0	1	2				
Nonlinearity (δ)	.2	.3	.4	.6	.8			
Total cases = 90								

Table 6.3-1. SUMMARY OF THE 192 SPECTRAL DISTANCE MEASURES STUDIED

number of results. What is done, instead, is to use this new data base of results to answer specific questions of interest about the utility of sample spectral distance measures and the optimality of the controlling parameters.

6.3.1 The Best Spectral Distance Measures

The first question of interest is what are the best spectral distance measures and how good are they. Table 6.3.1-1 gives a list of the five best spectral distance measures for CA, TSQ, and TBQ for ALL and WBD.

Several points should be noted here. First the best measure for the spectral distance measure overall distortions for CA uses the $|\ |^{.2}$ nonlinearity and uses neither energy weighting nor spectral weighting. The $|\ |^{.2}$ nonlinearity is very close to the log nonlinearity over much of its range, and indeed, two log measures are included in the top five.

The maximum correlation coefficient is -.6020, corresponding to a standard deviation of error of 7.86. This is not very good, and even though this is one of the better simple measures, it does not do very well. This is a general result and clearly indicates that composite measures are necessary if effective objective measures are to be designed.

The results over TSQ are similar, though slightly lower, than those for CA. Here, the log measures are consistently better than those using the $|\ |^{\delta}$ nonlinearity.

By comparison, the results for TBQ are very poor, with a maximum correlation of only .135. Note that these correlations are all positive, as would be expected. Since all the spectral distance measures explicitly measure signal distortion, it is not surprising that they do a poor job on background qualities.

	$\hat{\rho}$	$\hat{\sigma}_e$	Nonlinearly (δ)	Lp Norm (P)	Spectral Weighting (γ)	Energy Weighting (α)
CA (ALL)	.60	7.9	ll .2	2	0.0	0.0
	.60	7.9	ll .2	2	0.0	0.5
	.60	7.9	log	4	0.0	0.0
	.59	7.9	log	2	1.0	0.0
	.59	7.9	ll .2	4	0.0	0.0
CA(WBD)	.63	7.0	log	2	1.0	1.0
	.63	7.0	log	4	2.0	1.0
	.63	7.0	log	8	2.0	1.0
	.63	7.0	log	4	1.0	1.0
	.63	7.0	log	2	1.0	0.5
TSQ(ALL)	.57	8.8	log	2	1.0	2.0
	.57	8.8	log	2	1.0	1.0
	.57	8.8	log	1	1.0	2.0
	.57	8.8	log	1	2.0	1.0
	.57	8.8	log	1	1.0	2.0
TSQ(WBD)	.64	7.5	log	8	2.0	2.0
	.64	7.5	log	4	2.0	2.0
	.64	7.5	log	4	1.0	2.0
	.64	7.5	log	8	2.0	1.0
	.64	7.5	log	4	2.0	1.0
TBQ(ALL)	.14	7.2	linear	1	0.0	2.0
	.13	7.2	linear	2	0.0	2.0
	.13	7.2	linear	2	1.0	2.0
	.13	7.2	linear	1	1.0	2.0
	.13	7.2	linear	4	2.0	2.0
TBQ(WBD)	.23	6.2	ll .6	1	0.0	2.0
	.23	6.2	ll .4	1	0.0	2.0
	.23	6.2	ll .8	1	0.0	2.0
	.23	6.2	ll .6	1	0.0	1.0
	.22	6.2	ll .8	1	0.0	1.0

Table 6.3.1-1. Best Five Spectral Distance Measures for CA, TSQ, and TBQ Across ALL and WBD.

The correlations of all the measures over the WBD set show about a .03-.08 point improvement over the ALL set. This, of course, is a more realistic estimate of how these parameters would perform on true coding distortions. Here, the best result for CA is $\hat{\rho} = -.6345$ with $\hat{\sigma}_e = 7.086$ and for TSQ is $\hat{\rho} = .6427$ with $\hat{\sigma}_e = 7.67$, with the log nonlinearity always the best. The results for TBQ are once again very poor, though significantly better than for ALL.

6.3.2 The Effect of Energy Weighting

The effect of energy weighting was tested for spectral distance measures using the ALL and WBD data sets, for four groupings: all spectral distance measures; log spectral distance measures; linear spectral distance measures; and $|\ |^{\delta}$ spectral distance measures. The composite rank analyses for this test is shown in Table 6.3.2.

The results for energy weighting here are very clear. Energy weighting does not help. The ranking for α is 0 - .5 - 1 - 2, where 0 or no weighting is best. This is a very strong result for all the spectral distance classes. Note also that the only deviation from this strong (in fact, perfect) result occurs in the linear spectral distance case. However, linear spectral distance measures consistently perform poorly, so these deviations are of little interest.

6.3.3 The Effects of Spectral Weighting

The effects of weighting the spectral distance measure in the frequency domain by $|V(n,s,d,\theta)|^{\gamma}$ was tested for ALL and WBD for $\gamma = 0,1,$ and 2 across the same for groups of spectral distance measure used in section 6.3.3. The results of this study are shown in Table 6.3.3-1.

Group Tested		Energy Weighting Parameter (α)			
		0	.5	1	2
All spectral	0	--	1.04	1.02	1.00
distance	.5	$.4 \times 10^{-11}$(48)	--	1.00	1.00
measures	1	$.2 \times 10^{-12}$(48)	$.4 \times 10^{-14}$(48)	--	1.00
	2	$.4 \times 10^{-14}$(48)	$.4 \times 10^{-14}$(48)	$.4 \times 10^{-14}$(48)	--
		0	.5	1	2
Log spectral	0	--	1.00	1.00	1.00
distance	.5	10^{-6}(20)	--	1.00	1.00
measures	1	10^{-6}(20)	10^{-6}(20)	--	1.00
	2	10^{-6}(20)	10^{-6}(20)	10^{-6}(20)	--
		0	.5	1	2
Linear spectral	0	--	1.00	1.00	1.00
distance	.5	$.2 \times 10^{-5}$(16)	--	1.00	1.00
measures	1	$.2 \times 10^{-5}$(16)	$.2 \times 10^{-5}$(16)	--	1.00
	2	$.2 \times 10^{-5}$(16)	$.2 \times 10^{-5}$(16)	$.2 \times 10^{-5}$(16)	--
		0	.5	1	2
$\lvert\cdot\rvert$ spectral	0	--	1.67	1.08	1.00
distance	.5	.019(12)	--	1.00	1.00
measures	1	.003(12)	.002(12)	--	1.00
	2	.002(12)	.002(12)	.002(12)	--

Table 6.3.2. RANK TEST RESULTS FOR ENERGY WEIGHTING

Each frame was weighted by the energy in the undistorted speech frame to the α power.

Spectral Weighting Parameter (γ)

		0	1	2
All spectral	0	--	1.84	1.50
distance	1	$.5\text{x}10^{-4}$(32)	--	1.28
measures	2	.57(32)	.01(32)	--

Log spectral		0	1	2
distance	0	--	2.00	1.31
measures	1	$.15\text{x}10^{-4}$(16)	--	1.06
	2	.10(16)	$.2\text{x}10^{-3}$(16)	--

		0	1	2
Linear spectral	0	--	1.68	1.68
distance	1	.10(16)	--	1.50
measures	2	.10(16)	.59(16)	--

Table 6.3.3-1. RANK TEST RESULTS FOR SPECTRAL WEIGHTING BY $V(m,p,d,\theta)^{\gamma}$ FOR SPECTRAL DISTANCE MEASURES.

The consistent result here is that the second case, $\gamma=1$, is significantly better than $\gamma=0$ at both the .05 and .01 level, but is significantly better than $\gamma=2$ at only the .1 level. Once again, this result is weaker for the case of linear spectral distance measures. So the basic result is that $\gamma=1$ should be used, but this is a weak statement.

6.3.4 The Effects of L_p Averaging

The effects of L_p averaging in the frequency domain for p = 1,2,4, 8,10,12,14, and 16 were tested for ALL and WBD across the same spectral distances groups as in the last two tests. The results of the study are given in Table 6.3.4-1.

When viewed across all the spectral distance measures, the results are mixed, with p=1 best, but not significantly so. However, the individual results here show a very different picture. The linear spectral distance measures, the ranking is p = 1-2-4-8, where every result is significant at the .01 level. Since linear spectral distance does not perform well, this is not an interesting result. For the log spectral distance measure, the ranking is p = 4 - 8 - 2 -11 - 10 - 12 - 14 - 16, where the only non-significant results occur between the 4 and 8 levels. (Note that the lack of significance generally associated with the 10 - 12 - 14 - 16 levels is due to the lack of samples.) This is a very powerful and interestingresult since most researchers have used p=1 or p=2 in utilizing log spectral distance measures. These results clearly show that a value of p between 4 and 8 will work better.

The results for $|\ |^{\gamma}$spectral distance measure are mixed. This is clearly expected since this measure, in a sense, forms a bridge of nonlinearities between the linear ($\delta=1$) and the log ($\delta\approx.33$) nonlinearity.

		Lp NORM PARAMETER (p)							
		1	2	4	8	10	12	14	16
All Spectral Distance Measures	1	---	1.39	1.29	1.15	1.00	1.00	1.00	1.00
	2	.08(44)	---						
	4	.005(44)	$.1x10^{-4}$(44)	---	1.18	1.36	1.00	1.00	1.00
	8	$.3x10^{-5}$(44)	$.5x10^{-6}$(44)	$.8x10^{-8}$(44)	---				
	10	.06(4)	.06(4)	.06(4)	.06(4)	---	1.09	1.00	1.00
	12	.06(4)	.06(4)	.06(4)	.06(4)	.06(4)	---		
	14	.06(4)	.06(4)	.06(4)	.06(4)	.06(4)	.06(4)	---	1.00
	16	.06(4)	.06(4)	.06(4)	.06(4)	.06(4)	.06(4)	.06(4)	---
Log Spectral Distance Measures	1	---	2.00	2.00	1.58	1.00	1.00	1.00	1.00
	2	$.2x10^{-3}$(12)	---	1.67	1.5	1.00	1.00	1.00	1.00
	4	$.2x10^{-3}$(12)	.19(12)	---	1.33	1.00	1.00	1.00	1.00
	8	.38(12)	.61(12)	.19(12)	---	1.00	1.00	1.00	1.00
	10	.06(4)	.06(4)	.06(4)	.06(4)	---	1.00	1.00	1.00
	12	.06(4)	.06(4)	.06(4)	.06(4)	.06(4)	---	1.00	1.00
	14	.06(4)	.06(4)	.06(4)	.06(4)	.06(4)	.06(4)	---	1.00
	16	.06(4)	.60(4)	.06(4)	.06(4)	.06(4)	.06(4)	.06(4)	---

Table 6.3.4-1(a) RANK TEST RESULTS FOR L_p NORM FOR SPECTRAL DISTANCE MEASURES

		1	2	4	8
Linear Spectral Distance Measures	1	---	1.00	1.00	1.00
	2	$.2x10^{-3}(12)$	---	1.00	1.00
	4	$.2x10^{-3}(12)$	$.2x10^{-3}(12)$	---	1.00
	8	$.2x10^{-3}(12)$	$.2x10^{-3}(12)$	$.2x10^{-3}(16)$	---
$\|\|\delta$ Spectral	1	---	1.25	1.05	1.00
	2	.02(20)	---	1.00	1.00
	4	$.2x10^{-4}(20)$	$10^{-6}(20)$	---	1.00
	8	$10^{-6}(20)$	$10^{-6}(20)$	$10^{-6}(20)$	---

Table 6.3.4-1(b). RANK TEST RESULTS FOR Lp NORM FOR SPECTRAL DISTANCE MEASURES

6.3.5 The Effect of the Pointwise Nonlinearity

In the study of the effects of the pointwise nonlinearities, the cases considered were $|\ |^{\delta}$ for δ = 1,.2,.3,.4,.6, and .8 plus log. The results are shown in Table 6.3.5-1.

The basic result here is that the ranking is .2 - log - .3 - .4 - .6 -.8 - 1 where there is no significant difference between the δ=.2 case and the log, but all other differences are significant at the .01 level. This means that (1) a nonlinearity should be used (linear was ranked lowest), and (2) the log, $||^{.2}$, and $||^{.3}$ give very similar results. These three functions are indeed very similar over most of their ranges.

6.3.6 The Effects of Other Subjective Measures

Table 6.3.6-1 shows the maximum correlation value found for spectral distance measures over ALL and WBD for nine different isometric subjective quality measures available from the DAM; Composite Acceptability, CA; Total System Quality, TSQ; Total Background Quality, TBQ; parametric Pleasantness, PP; Parametric Intelligibility, PI; Parametric Acceptability, PA; raw Pleasantness, P; raw Intelligibility, I; and raw acceptability, A. The maximum values are given here, since they were fairly representative of the overall results for the entire subjective parameter.

Several things are noteworthy here. First, note, as before, TBQ is not tracked well by the objective measures. Second, note that the behavior is similar over all the measures, but with intelligibility measures (PI and I) being tracked better than the rest. The worst tracking of a system quality was for pleasantness (PP and P), with acceptability showing intermediate behavior.

NONLINEARITY PARAMETER

δ

	log	.2	.3	.4	.6	.8	1
log	---	1.81	1.38	1.25	1.00	1.00	1.00
.2	.01(16)	---	1.25	1.00	1.00	1.00	1.00
.3	.22(16)	.04(16)	---	1.00	1.00	1.00	1.00
.4	$.15\times10^{-4}$(16)	$.15\times10^{-4}$(16)	$.15\times10^{-4}$(16)	---	1.00	1.00	1.00
.6	$.15\times10^{-4}$(16)	$.15\times10^{-4}$(16)	$.15\times10^{-4}$(16)	$.15\times10^{-4}$(16)	---	1.00	1.00
.8	$.15\times10^{-4}$(16)	$.15\times10^{-4}$(16)	$.15\times10^{-4}$(16)	$.15\times10^{-4}$(16)	$.15\times10^{-4}$(16)	---	1.00
1	$.9\times10^{-12}$(40)	10^{-6}(20)	10^{-6}(20)	$.15\times10^{-4}$(20)	$.15\times10^{-4}$(20)	$.15\times10^{-4}$(20)	---

Table 6.3.5-1. PAIRWISE RANK TEST FOR δ ON THE $\int^{\delta}$ NONLINEARITY PLUS THE LOG NONLINEARITY

DISTORTION SET		CA	TSQ	TBQ	PP	PI	PA	P	I	A
ALL	$\hat{\rho}$	-.60	-.57	.14	-.53	-.64	-.53	-.51	-.66	-.61
	$\hat{\sigma}_e$	7.8	8.8	6.0	8.5	8.2	8.3	7.8	6.8	8.3
WBD	$\hat{\rho}$	-.63	-.64	.23	-.50	-.69	-.62	-.47	-.71	-.64
	$\hat{\sigma}_e$	7.0	7.7	6.0	7.3	8.2	6.9	6.5	6.9	7.0

Table 6.3.6-1. MAXIMUM CORRELATION OVER ALL SPECTRAL DISTANCE MEASURES FOR DIFFERENT SUBJECTIVE MEASURES

6.3.7 The Effects of Different Distorted Data Bases

Table 6.3.7-1 shows the results for the best spectral distance measures for CA, TSQ, and TBQ over all the distorted data base subsets (see section 6.2.2). There are several surprising features of this data. First, the performance of the individual measures of many of the subsets is surprisingly uniform. This suggests there is only a slight gain to be expected from these measures if there is a preclassification step in the analysis. Another interesting result is the measures performance on wide band noise vs. narrow band noise. It does outstandingly on narrow band noise, and not very well on wide band noise. This is probably due to the fact that no energy measurement is included in these spectral distance measures.

6.3.8 The Effects of Nonlinear Regression Analysis

In order to study the effects of using higher order regression analysis, the CA, TSQ and TBQ subjective measures were tested for third and sixth degree regression analysis across the ALL and WBD distorted data base. Table 6.3.8-1 gives a compilation of these results for the best measures observed. In both the CA and TSQ cases, it would appear that one obtains remarkable improvements by going to higher order regression analysis. In the most remarkable case, sixth order WBD across CA gives a correlation of .98 and a $\hat{\sigma}_e$ on only 1.7. One must be very careful in analyzing these results. Clearly, the more parameters in the nonlinear approximation which are set optimally, the better the results will be. This, of course, is a mathematical certainty. As we allow larger higher order regressions, at some point we begin to track the noise in the system. In this sense, the numbers presented here should be considered approximate

DISTORTION SET		CA	TSQ	TBQ
ALL	$\hat{\rho}$	-.60	-.57	.13
	$\hat{\sigma}_e$	7.8	8.8	7.2
WBD	$\hat{\rho}$	-.63	-.64	.23
	$\hat{\sigma}_e$	7.0	7.7	6.0
CODE	$\hat{\rho}$	-.65	-.64	-.30
	$\hat{\sigma}_e$	6.1	6.8	6.6
CON	$\hat{\rho}$	-.63	-.64	.21
	$\hat{\sigma}_e$	8.3	8.6	7.5
WBN	$\hat{\rho}$	-.58	-.57	-.29
	$\hat{\sigma}_e$	6.2	7.1	6.5
NBN	$\hat{\rho}$	-.92	-.83	-.87
	$\hat{\sigma}_e$	3.6	3.4	3.8
BD	$\hat{\rho}$	-.65	-.67	-.50
	$\hat{\sigma}_e$	5.6	7.6	4.1
PD	$\hat{\rho}$	-.67	-.67	-.30
	$\hat{\sigma}_e$	6.0	6.2	7.4

Table 6.3.7-1. MAXIMUM CORRELATION VALUES FOR SPECTRAL DISTANCE MEASURES FOR CA, TSQ, AND TBQ OVER THE DIFFERENT SUBSETS OF THE DISTORTED DATA BASE.

CA

		1st Order	3rd Order	6th Order
ALL	$\hat{\rho}$	.60	.69	.80
	$\hat{\sigma}_e$	7.8	7.1	5.8
WBD	$\hat{\rho}$	.63	.73	.98
	$\hat{\sigma}_e$	7.0	6.1	1.7

TSQ

		1st Order	3rd Order	6th Order
ALL	$\hat{\rho}$	.57	.64	.75
	$\hat{\sigma}_e$	8.8	8.2	7.0
WBD	$\hat{\rho}$	.64	.70	.88
	$\hat{\sigma}_e$	7.7	7.1	4.61

TBQ

		1st Order	3rd Order	6th Order
ALL	$\hat{\rho}$	.14	.28	.44
	$\hat{\sigma}_e$	6.0	6.9	6.4
WBD	$\hat{\rho}$	.23	.42	.84
	$\hat{\sigma}_e$	6.0	5.5	3.3

Table 6.3.8-1. THE EFFECTS OF NON-LINEAR REGRESSION ANALYSIS ON SPECTRAL DISTANCE MEASURES. ONLY MAXIMUM RESULTS ARE SHOWN.

upper limits on the performance of measures based on higher order regression models.

In spite of the above warning, these results are very promising. Certainly, the third order effects are probably attainable in a real system. Because of the apparent improvement attainable from these polynomial pointwise nonlinearities, it would be of great interest to investigate other forms of this nonlinearity.

For the case of TBQ, the improvements are equally remarkable. However, as before, the spectral distance measures are relatively ineffective at predicting the subjective background quality.

6.4 Simple Noise Measures

The simple noise measures studied, as described in Section 3.3.3, include both the ordinary SNR and the "short time" SNR. In this study, only four measures were studied: the ordinary SNR and the short time SNR with δ = .5, 1, and 2. In all the short time studies, the frame interval was taken to be 256 points. Previous researchers [6.1] have indicated that this measure is relatively insensitive to the frame interval. This measure, of course, is only meaningful over the waveform coders and those controlled distortions which can be thought of as being additive noise. Hence, these measures were only tested across WFC (waveform coders) and ND (noise distortions). Table 6.4-1 shows the results of these experiments.

The first obvious point here is that the traditional SNR is not a very good objective measure. By comparison, all forms of the short time SNR always perform better. The performances of all the measures are comparable over the WFC and ND distortion sets, and this should be a good estimate of their expected performance in real coding tests. The best

	WFC		ND	
SNR	.24	8.8	.31	8.8
Short time SNR (δ=.5)	.76	5.6	.77	5.9
Short time SNR (δ=1)	.77	5.7	.78	6.0
Short time SNR (δ=2)	.75	5.5	.77	5.9

Table 6.4-1. RESULTS FOR SNR AND SHORT TIME SNR FOR CA ACROSS WFC AND ND

value of δ was found to be 1, though the differences between the three values were small.

The clear point here is that the short time SNR is clearly superior to the traditional SNR, and should replace this measure whenever possible.

6.5 The Parametric Distance Measures

The parametric distance measures, as discussed in Section 3.3.2, can be divided into seven classes; feedback coefficient distance measures; log feedback coefficient distance measures; PARCOR distance measures; log PARCOR distance measures; area ratio distance measures; log area ratio distance measures; and the energy ratio measure. In the experimental study, the first six measures were studied as a group, while the energy ratio measure, because of its wide use, was studied separately. In all, 38 forms of the energy ratio measure and 72 forms of the other measures were studied. The overall experimental philosophy was the same for these measures as for the spectral distance measures, and a similar set of experiments were conducted. These are isometric measures, so, as before, the ALL and WBD distortion subsets are used to predict their effectiveness.

Within each of the seven classes of parametric distance measure, the particular distance measure may be described by two conditions: the value of p in the L_p norm; and the energy weighting parameter, α. In terms of these parameters, Table 6.5-1 describes the measures tested for each of the seven classes.

6.5.1 The Best Parametric Distance Measures

The best parametric distance tested was found to be the L_1 log area ratio measure without energy weighting. This measure has a correlation coefficient of -.62 for CA across ALL, and a correlation coefficient of

	Lp Norm (P)	Energy Weighting(α)	Total
Linear feedback	1, 2, 4	0, 1, 2	9
Log feedback	1, 2, 4	0, 1, 2	9
Linear PARCOR	1, 2, 4	0, 1, 2	9
Log PARCOR	1, 2, 4	0, 1, 2	9
Linear area ratio	1, 2, 4	0, 1, 2	9
Log area ratio	1, 2, 4	0, 1, 2	9
Energy ratio measure	.25,.5,1,2,4,8	0,.25,.5,1,2,4,8	38

Table 6.5-1. SUMMARY OF PARAMETERS FOR PARAMETRIC DISTANCE MEASURES

-.65 for CA across WBD. This is a very important result, for it says that this parametric distance measure performed better than any of the spectral distance measures. Since this measure is an order of magnitude more compact to compute, this is a very important result.

Tables 6.5.1-1 through 6.5.1-7 give the best six measures for each of the seven categories for CA across ALL and WBC. The only two of these measures which show any promise are the log area ratio distance measure and the energy ratio distance measure. The results for these two measures will be presented in more detail.

6.5.2 The Log Area Ratio Measure

The results for the log area ratio measure tests are summarized in Table 6.5.2-1, which gives the results of all the log area ratio measures studied for CA, TSQ, and TBQ across CA and WBD. In each case, the log area ratio measure performs comparable to but better than the corresponding spectral distance measure. Like the spectral distance measure, performance was relatively poor for TBQ.

Table 6.5.2-2 shows the maximum results for the log area ratios across the other distortion subsets for CA. Here, once again, the results are comparable to but better than those from the spectral distance measures.

Table 6.5.2-3 shows the effects of third order and sixth order nonlinear regression. Improvements here are also comparable to those from spectral distance measures.

Examination of the data also shows other similarities to the spectral distance results. For the log area ratio, no energy weighting is best, followed by energy weighting to the first, then second power.

CA (ALL)

$\hat{\rho}$	$\hat{\sigma}_e$	p	α
.06	9.8	1	0
.06	9.8	2	0
.04	9.8	1	1
.03	9.8	2	1
.03	9.8	1	2
.03	9.8	2	2

CA (WBD)

$\hat{\rho}$	$\hat{\sigma}_e$	p	α
.14	8.9	2	0
.14	8.9	1	0
.12	8.9	2	1
.12	8.9	1	1
.11	8.9	2	2
.08	8.9	1	2

Table 6.5.1-1. BEST SIX RESULTS FOR LINEAR FEEDBACK PARAMETRIC DISTANCE MEASURE

CA (ALL)

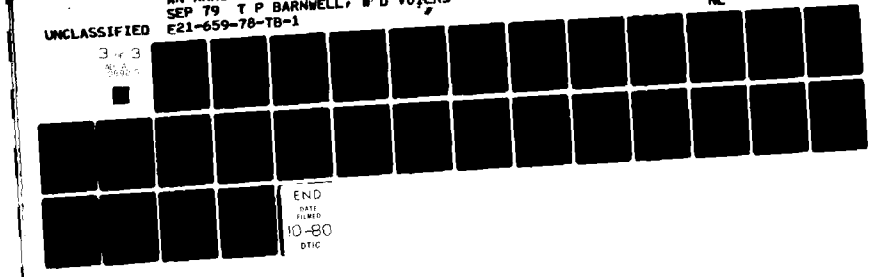
$\hat{\rho}$	$\hat{\sigma}_e$	p	α
.11	9.8	1	0
.10	9.8	2	0
.05	9.8	1	1
.05	9.8	2	1
.04	9.8	1	2
.04	9.8	2	2

CA (WBD)

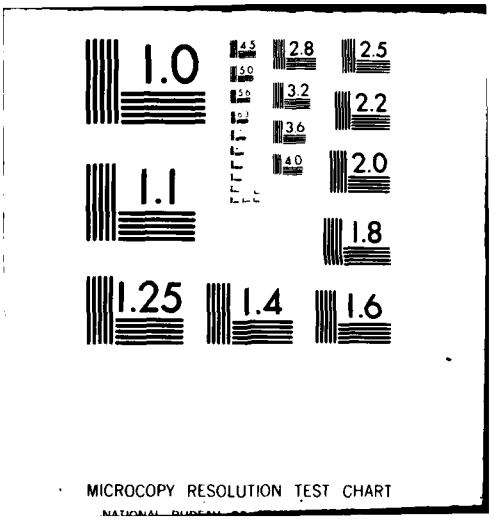
$\hat{\rho}$	$\hat{\sigma}_e$	p	α
.32	8.5	2	0
.31	8.6	1	0
.29	8.6	2	1
.28	8.6	1	1
.26	8.6	2	2
.25	8.7	1	2

Table 6.5.1-2. BEST SIX RESULTS FOR LOG PARCOR PARAMETER DISTANCE MEASURE

AD-A089 210 GEORGIA INST OF TECH ATLANTA SCHOOL OF ELECTRICAL EN--ETC F/G 17/2
AN ANALYSIS OF OBJECTIVE MEASURES FOR USER ACCEPTANCE OF VOICE --ETC(U)
SEP 79 T P BARNWELL, W D VOIERS DCA100-78-C-0003
UNCLASSIFIED E21-659-78-TB-1 NL



END
DATE FILMED
10-80
DTIC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
MICROCOPY RESOLUTION TEST CHART

CA (ALL)

$\hat{\rho}$	$\hat{\sigma}_e$	p	α
.11	9.8	1	0
.11	9.8	2	0
.06	9.8	1	1
.06	9.8	2	1
.05	9.8	1	2
.05	9.8	2	2

$\hat{\rho}$	$\hat{\sigma}_e$	p	α
.33	8.5	2	0
.32	8.5	1	0
.31	8.5	2	1
.30	8.6	1	1
.28	8.6	2	2
.27	8.6	1	2

Table 6.5.1-3. BEST SIX RESULTS FOR LOG FEEDBACK COEFFICIENT PARAMETRIC DISTANCE MEASURE

CA (ALL)

$\hat{\rho}$	$\hat{\sigma}_e$	p	α
.24	9.6	1	0
.22	9.6	1	1
.21	9.6	1	2
.20	9.6	2	0
.19	9.7	2	1
.18	9.7	2	2

CA (WBD)

$\hat{\rho}$	$\hat{\sigma}_e$	p	α
.32	8.5	1	0
.30	8.6	2	0
.28	8.6	1	1
.28	8.6	1	2
.27	8.6	2	1
.26	8.7	2	2

Table 6.5.1-4. BEST SIX RESULTS FOR LINEAR AREA RATIO PARAMETRIC DISTANCE MEASURE

CA (ALL)

$\hat{\rho}$	$\hat{\sigma}_e$	p	α
.46	8.7	1	0
.30	9.3	2	0
.29	9.4	1	1
.21	9.6	1	2
.16	9.7	2	1
.12	9.8	2	2

CA (WBC)

$\hat{\rho}$	$\hat{\sigma}_e$	p	α
.43	8.1	1	0
.31	8.5	2	0
.30	8.6	1	1
.24	8.7	1	2
.21	8.8	2	1
.18	8.8	2	2

Table 6.5.1-5. SIX BEST RESULTS FOR THE LINEAR PARCOR PARAMETRIC DISTANCE MEASURE

CA (ALL)

$\hat{\rho}$	$\hat{\sigma}_e$	p	α
.62	7.7	1	0
.62	7.7	2	0
.62	7.8	1	1
.61	7.8	2	1
.60	7.9	1	2
.59	7.9	2	2

CA (WBC)

$\hat{\rho}$	$\hat{\sigma}_e$	p	α
.65	6.8	1	0
.64	6.9	2	0
.64	6.9	1	1
.64	6.9	2	1
.64	6.9	1	2
.62	7.0	2	2

Table 6.5.1-6. BEST SIX RESULTS FOR LOG AREA RATIO PARAMETRIC DISTANCE

CA (ALL)

$\hat{\rho}$	$\hat{\sigma}_e$	p	Energy Weighting(α)
.60	7.9	.25	0.0
.58	8.0	.5	0.0
.53	8.3	.5	.25
.51	8.5	2.5	.25
.49	8.6	1.0	1.0
.49	8.6	1.0	.50

CA (WBD)

$\hat{\rho}$	$\hat{\sigma}_e$	p	Energy Weighting(α)
.65	6.8	.25	0.0
.63	7.0	.50	0.0
.62	7.0	.50	1.0
.62	7.0	.25	.25
.61	7.1	.50	.50
.61	7.1	.25	.50

Table 6.5.1-7. BEST SIX RESULTS FOR THE ENERGY RATIO PARAMETRIC DISTANCE MEASURE

	$\hat{\rho}$	$\hat{\sigma}_e$	L_p NORM p	ENERGY WEIGHTING (α)
CA (ALL)	.62	7.7	1	0
	.62	7.7	2	0
	.62	7.8	1	1
	.61	7.8	2	1
	.60	7.9	1	2
	.59	7.9	2	2
CA (WBD)	.65	6.8	2	1.0
	.64	6.9	2	2.0
	.64	6.9	1	1.0
	.64	6.9	2	0.0
	.64	6.9	1	2.0
	.62	7.0	1	0.0
TSQ (ALL)	.58	8.7	1	2.0
	.58	8.8	1	1.0
	.57	8.8	2	1.0
	.57	8.8	2	0.0
	.54	9.0	1	0.0
	.52	9.1	2	0.0
TSQ (WBD)	.62	7.0	2	1.0
	.61	7.1	2	2.0
	.61	7.1	1	2.0
	.60	7.2	1	1.0
	.59	7.2	2	0.0
	.58	7.3	1	0.0
TBQ (ALL)	.11	7.2	1	0.0
	.11	7.2	2	0.0
	.03	7.2	1	1.0
	.2	7.2	2	1.0
	.006	7.2	2	2.0
	.0006	7.2	1	2.0
TBQ (WBD)	.15	6.1	2	2.0
	.15	6.1	1	2.0
	.14	6.1	2	1.0
	.13	6.1	1	1.0
	.10	6.1	2	0.0
	.10	6.1	1	0.0

Table 6.5.2-1. TOTAL RESULTS FOR LOG AREA RATIO PARAMETRIC MEASURE FOR CA, TSQ, AND TBQ FOR ALL AND WBD

Distortion Subset	$\hat{\rho}$	$\hat{\sigma}_e$
ALL	.62	7.7
WBD	.65	6.8
WFC	.64	6.9
CODE	.62	6.2
CON	.65	8.2
WBN	.40	7.0
NBN	.91	3.8
BD	.58	6.0
PD	.53	6.9

Table 6.5.2-2. THE MAXIMUM VALUES FOR CA FOR THE LOG AREA RATIO MEASURE ACROSS DIFFERENT DISTORTION SUBSETS

ANALYSIS ORDER

	1st Order		3rd Order		6th Order	
	$\hat{\rho}$	$\hat{\sigma}_e$	$\hat{\rho}$	$\hat{\sigma}_e$	$\hat{\rho}$	$\hat{\sigma}_e$
CA (ALL)	.62	7.7	.64	7.5	.69	7.1
CA (WBD)	.65	6.8	.66	6.7	.79	5.5
TSQ (ALL)	.58	8.7	.59	8.7	.72	7.4
TSQ (WBD)	.62	7.0	.63	7.0	.72	6.1
TBQ (ALL)	.11	7.2	.24	7.0	.42	6.6
TBQ (WBD)	.15	6.1	.35	8.4	.94	3.1

Table 6.5.2-3. THE EFFECTS OF HIGHER ORDER REGRESSION ANALYSIS ON THE LOG AREA RATIO DISTANCE MEASURE

6.5.3 The Energy Ratio Distance Measure

The results for the energy ratio distance measure are summarized in Tables 6.5.3-1, 6.5.3-2, and 6.5.3-3. The first table gives maximum results for CA, TSQ, and TBQ over CA and WBD. The second table gives maximum results for CA over the other distortion subsets. The third table shows the results of nonlinear regression analysis.

The energy ratio distance measure does quite well in all tests, but it is not able to quite match the performance of either the log area ratio measure or the best spectral distance measure. The general performance of all three of these measures is very similar, with the energy ratio measures being the poorest of the three. This is probably because these measures are measuring very similar features of the speech distortions.

6.6 Frequency Variant Measures

There are two basic classes of frequency variant measures studied as part of this research: frequency variant spectral distance measures; and frequency variant noise measurement. For both cases, the frequency range 200-3200 Hz is divided into six bands, as shown in Table 6.6-1. The individual measures for each of the bands is then computed, and the overall objective measure is formed as an optimally weighted sum of the subband results.

6.6.1 The Frequency Variant Spectral Distance Measures

The parameters controlling the frequency variant spectral distance measures are the same as those controlling the spectral distance measures. These include four conditions. First, the distance measure may be between linear spectra, log spectra, or spectra taken to the δ power. Second, the

	CA		WBD	
	$\hat{\rho}$	$\hat{\sigma}_e$	$\hat{\rho}$	$\hat{\sigma}_e$
CA	.59	7.9	.65	6.9
TSQ	.54	9.0	.62	7.0
TBQ	.12	7.1	.24	6.8

Table 6.5.3-1. MAXIMUM RESULTS FROM THE ENERGY RATIO DISTANCE MEASURE

Distortion Subset	$\hat{\rho}$	$\hat{\sigma}_e$
ALL	.59	7.9
WBD	.61	6.9
WFC	.58	6.7
CON	.59	8.7
CODE	.53	6.7
WBN	.47	6.7
NBN	.80	5.5
BD	.60	5.9
PD	.57	6.7

Table 6.5.3-2. THE MAXIMUM VALUE OF CA FOR THE ENERGY RATIO MEASURE ACROSS DIFFERENT DISTORTION SUBSET

ANALYSIS CODE	1st Order		3rd Order		6th Order	
	$\hat{\rho}$	$\hat{\sigma}_e$	$\hat{\rho}$	$\hat{\sigma}_e$	$\hat{\rho}$	$\hat{\sigma}_e$
CA(ALL)	.59	7.9	.64	7.6	.64	7.5
CA(WBD)	.65	6.9	.66	6.7	.68	6.6
TSQ(ALL)	.54	9.0	.38	8.8	.60	8.6
TBQ(ALL)	.62	7.0	.63	7.0	.65	6.8
TBQ(ALL)	.12	7.1	.24	7.0	.69	5.2
TBQ(WBD)	.24	6.8	.30	6.7	.36	6.4

Table 6.5.3-3. THE EFFECTS OF HIGHER ORDER REGRESSION ANALYSIS ON THE ENERGY RATIO DISTANCE MEASURE

BAND NUMBER	RANGE (HZ)
1	200-400
2	400-800
3	800-1300
4	1300-1900
5	1900-2600
6	2600-3400

Table 6.6-1. FREQUENCY BANDS USED FOR THE FREQUENCY VARIANT OBJECTIVE MEASURES

distance measure may be frequency weighted by the energy spectrum, $V(n,s,d,\theta)^{\gamma}$. Third the measure may be time weighted by the energy of the frame taken to the α power. And finally, of course, an Lp norm is evolved, and the value of p is an important parameter. In all, a total of 96 variations of these measures were studied. These measures are summarized in Table 6.6.1-1.

Table 6.6.1-2 shows the results for the five best log spectral distance measures. As can be seen, the use of frequency weighting improves the spectral distance results by about .1 points in the correlation measure. Also, it was found that the same log spectral distance measures which did well in the non-frequency-variant cases did well in the frequency variant cases as well.

Table 6.6.1-3 shows the results for the five best linear spectral distance measures. Here, the improvement from the non-frequency-variant case is remarkable. Not only is the frequency variant linear spectral distance measure better than the non-frequency-variant case, it is better than the log measure also. This is an important result.

An important point about these frequency variant measures is that they are "tunable" for parametric as well as isometric subjective quality measures. Hence, correlation analyses were performed for the parametric subjective categories of SF,SH,SD,SL,SI,SN,BN,BB,BF, and BR across ALL and WBC. Table 6.6.1-4 shows some results from that study.

Qualitatively, these results are relatively easy to understand. Basically, the frequency variant spectral distance measures did well on frequency variant parametric subjective measures (SF,SH,SL,SN,BN, and BB) and poorly on the non-spectrally-related subjective measures (SP,SI,BF,

Linear Spectral Distance Measure	
Spectral weighting parameter (γ)	0, .5, 1, 2
Energy weighing parameter (α)	0, 1, 2
Lp Norm (p)	1, 2, 4, 8
TOTAL	48
Log Spectral Distance Measure	
Spectral weighting parameter (γ)	0, .5, 1, 2
Energy weighting parameter (α)	0, 1, 2
Lp Norm (p)	1, 2, 4, 8
TOTAL	48

Table 6.6.1-1. SUMMARY OF 96 FREQUENCY VARIANT SPECTRAL DISTANCE MEASURES TESTED

LOG FREQUENCY VARIANT SPECTRAL DISTANCE MEASURES

Condition			Spectral Weighting (γ)	Energy Weighting (α)	Lp Norm (p)
CA (ALL)	.68	7.2	1.0	0.0	4
	.68	7.2	1.0	0.0	2
	.68	7.2	2.0	0.0	8
	.67	7.3	1.0	0.0	1
	.67	7.3	1.0	0.0	8
CA (WBD)	.72	6.2	1.0	0.0	2
	.72	6.2	1.0	0.0	4
	.71	6.3	1.0	0.0	8
	.71	6.3	1.0	0.0	4
	.70	6.4	0.5	0.0	2
TSQ (ALL)	.61	8.5	1.0	1.0	2
	.61	8.5	2.0	1.0	4
	.61	8.5	2.0	1.0	8
	.60	8.6	1.0	1.0	2
	.60	8.6	1.0	0.0	4
TSQ (WBD)	.64	7.7	2.0	2.0	8
	.64	7.7	2.0	2.0	4
	.64	7.7	2.0	1.0	4
	.64	7.7	1.0	2.0	8
	.64	7.8	1.0	2.0	4
TBQ (ALL)	.23	6.0	2.0	0.0	1
	.23	6.0	2.0	0.0	2
	.22	6.0	2.0	1.0	2
	.22	6.0	2.0	1.0	1
	.22	6.0	2.0	2.0	4
TBQ (WBD)	.35	5.8	2.0	0.0	1
	.34	5.8	2.0	0.0	1
	.34	5.8	1.0	0.0	1
	.33	5.8	2.0	0.0	2
	.32	5.8	0.5	0.0	1

Table 6.6.1-2. BEST FIVE SYSTEMS FOR EACH CATEGORY FOR LOG FREQUENCY VARIANT SPECTRAL DISTANCE MEASURES

LINEAR FREQUENCY VARIANT SPECTRAL DISTANCE MEASURE

Condition			Spectral Weighting (γ)	Energy Weighting (α)	Lp Norm (p)
CA (ALL)	.68	7.2	0.0	2	1
	.68	7.2	0.0	2	2
	.68	7.2	0.5	2	1
	.68	7.2	0.0	1	1
	.68	7.2	0.0	2	4
CA (WBD)	.72	6.2	0.0	2	1
	.71	6.3	0.0	2	2
	.70	6.4	0.0	1	1
	.70	6.4	0.0	1	2
	.70	6.4	0.5	2	1
TSQ (ALL)	.61	8.5	0.5	2	1
	.61	8.5	1.0	2	1
	.61	8.5	0.5	2	2
	.61	8.5	0.5	1	1
	.61	8.5	1.0	1	1
TSQ (WBD)	.68	7.3	0.0	2	1
	.67	7.4	0.5	2	1
	.67	7.4	0.0	2	2
	.67	7.4	0.5	2	2
	.67	7.4	0.5	1	1
TBQ (ALL)	.24	7.0	0.0	2	1
	.24	7.0	0.0	1	1
	.23	7.0	0.0	2	2
	.23	7.0	0.0	1	2
	.22	7.1	0.0	2	4
TBQ (WBD)	.38	5.7	0.0	0	2
	.38	5.7	0.0	1	2
	.38	5.7	0.0	2	2
	.38	5.7	0.0	2	4
	.38	5.7	0.0	1	1

Table 6.6.1-3. BEST FIVE SYSTEMS FOR EACH CATEGORY FOR LINEAR FREQUENCY VARIANT SPECTRAL DISTANCE MEASURES

Parametric Subjective Measure	ALL $\hat{\rho}$	ALL $\hat{\sigma}_e$	WBD $\hat{\rho}$	WBD $\hat{\sigma}_e$
SF	.61	3.8	.74	4.2
SH	.63	3.9	.73	3.9
SD	.42	6.3	.26	6.7
SL	.72	3.6	.81	3.6
SI	.17	5.6	.19	7.9
SN	.45	3.8	.55	4.2
BN	.48	5.1	.23	4.0
BB	.43	4.0	.26	3.5
BF	.18	6.5	.38	5.4
BR	.27	2.8	.21	1.8

Table 6.6.1-4. SAMPLE OF RESULTS FOR FREQUENCY VARIANT SPECTRAL DISTANCE MEASURES USED FOR PREDICTING PARAMETRIC SUBJECTIVE RESULTS

and BR). The performance on several of the measures (SF,SH,SL) can be said to be very good, while the performance on the others is moderate or poor.

6.6.2 Frequency Variant Noise Measurements

The frequency variant noise measures studied include both the frequency variant form of the ordinary SNR and the frequency variant form of the short time SNR. Only the one form of the frequency variant SNR was tested. However, 49 versions of the short time frequency variant SNR were tested. These different measures are characterized by two parameters. The first parameter, the energy weighting parameter α, controls the time domain weighting by the energy of the original speech. The second, δ, controls the power to which the log of the measure is taken (see 3.4.2). In terms of these parameters, the 49 cases studied are shown in Table 6.6.2-1.

Of course these noise measures, like all noise measures, cannot be used across the whole distorted data base. Hence, these tests were only run across WFC, WBN, NBN, BD, and PD. The most important of these is WFC (waveform coders), since it represents an estimate of the measures' performance in a real coding environment.

Table 6.6.2-2 shows the results for WFC. The first noteworthy point is that these are outstanding results, with the best measure having a correlation coefficient of .93 and a $\hat{\sigma}_e$ of only 3.28. Note also that this is not an isolated measure, but that several forms of the measure come close to this performance.

In order to test the best values for the various parameters, a rank order study was done on both α and δ. The results of these studies are shown in Tables 6.6.2-3 and 6.6.2-4. As can be seen, the ranking for δ is

Banded Short Time SNR

Energy Weighting (α)	0,.25,.5,1,2,4,8
Power of log (δ)	0,.25,.5,1,2,4,8
TOTAL	49

Table 6.6.2-1. SUMMARY OF 49 SHORT TIME BANDED *SIGNAL-TO-NOISE* RATIO MEASURE

	$\hat{\rho}$	$\hat{\sigma}_e$	Energy Weighting α	Power Parameter δ
CA (WFC)	.93	3.3	0.0	.25
	.93	3.3	0.0	.50
	.93	3.3	.25	.25
	.93	3.3	.25	.50
	.93	3.3	.50	.25
TSQ (WFC)	.81	3.6	0.0	.25
	.81	3.6	0.0	.50
	.81	3.6	.25	.25
	.81	3.6	.25	.50
	.81	3.6	0.0	1.00
TBQ (WFC)	.93	2.9	0.0	.25
	.93	2.9	.25	.25
	.93	2.9	0.0	.50
	.93	2.9	.25	.50
	.93	3.0	.50	.25

Table 6.6.2-2. BEST FIVE RESULTS FOR BANDED SHORT TIME SNR MEASURE ACROSS WFC

Energy Weighting Parameter (α)

	0	.25	.50	1.0	2.0	4.0	8.0
0	---	1.0	1.0	1.0	1.0	1.0	1.0
.25	.008(7)	---	1.0	1.0	1.0	1.0	1.0
.5	.008(7)	.008(7)	---	1.0	1.0	1.0	1.0
1.0	.008(7)	.008(7)	.008(7)	---	1.0	1.0	1.0
2.0	.008(7)	.008(7)	.008(7)	.008(7)	---	1.0	1.0
4.0	.008(7)	.008(7)	.008(7)	.008(7)	.008(7)	---	1.0
8.0	.008(7)	.008(7)	.008(7)	.008(7)	.008(7)	.008(7)	---

Table 6.6.2-3. RESULTS OF THE PAIRWISE RANKING TEST FOR THE ENERGY WEIGHTING PARAMETER, α, FOR THE SHORT TIME SIGNAL-TO-NOISE RATIO

	Power Parameter δ					
	.25	.5	1.0	2.0	4.0	8.0
.25	---	1.14	1.0	1.0	1.0	1.0
.5	.06 (7)	---	1.0	1.0	1.0	1.0
1.0	.008(7)	.008(7)	---	1.0	1.0	1.0
2.0	.008(7)	.008(7)	.008(7)	---	1.0	1.0
4.0	.008(7)	.008(7)	.008(7)	.008(7)	---	1.0
8.0	.008(7)	.008(7)	.008(7)	.008(7)	.008(7)	---

Table 6.6.2-4. RESULTS OF THE PAIRWISE RANKING TEST FOR THE POWER PARAMETER FOR THE BANDED SHORT TIME SIGNAL-TO-NOISE RATIO

.25-.5-1-2-4-8, with .25 and .5 giving similar results. The ranking for α is 0-.25-.5-1-2-4-8. So, as before, the best energy weighting is no energy weighting.

6.7 The Composite Distance Measures

The composite distance measures studied in this research were always taken to be linear sums of up to six of the simple or frequency variant measures already discussed. Basically, there were two types of composite measures studied: measures without preclassification and measures with preclassification. In the measures without preclassification, exactly the same composite objective measure was applied to all the distortions under study. In the measures with preclassification, each of the distortions was assigned a class, and a different composite measure was applied to each class. The preclassification technique was not extensively explored in this study, but was only used to differentiate the spectral coders, such as vocoders, from those coders which could be considered as signal plus noise.

The composite measures were used in two ways in this study. The first use was to determine if different single measures were really measuring the same quantity or were measuring some different quantity. If they measure the same quantity, then the correlation coefficient based on their composite measure show only slight improvement. If they measure a different quantity, then the correlation coefficient show more improvement.

The second use for the composite measure was to search for a reasonable measure to be used in an objective quality testing system which attempts to predict the subjective results from the objective results. Two points should be made about this study. First, since the optimization of

the composite measure involves the setting of certain of the parameters based on the data, the results found here are limits on the performance of these measures, and other tests need to be made concerning their robustness Second, the composite measure technique used here is essentially a "bulk" technique which allows the automated study of a number of combinations rather easily. It is undoubtedly true that some additional gain might be obtained from studying the measures "by hand", using interactive graphics, and making appropriate pragmatic changes in the definitions of the objective measures.

6.7.1 The Composite Measure Used to Measure Mutual Information

In this part of the study, a large number of six wide composite measures were designed to find to what extent the correlation coefficient could be improved by combining the results of specific groups. For example, composite measures were made from all log spectral distance measures. This would answer the question of whether all the log spectral distance measures really contained similar information, or if some contained different information. Similarly, composite measures between log spectral distance measures and log area ratio measures would determine if they measured different information. By no means are these tests all inclusive, but they do represent a reasonable sampling of the effects. Table 6.7.1-1 gives a summary of the maximum results for the classes studied for CA across WBC and, where appropriate, WFC.

The results here can be summarized as follows. First, from line 1, all the log spectral distances contain similar information. This is true to a lesser extent (line 2) of all classes of spectral distance measures. From line 3, the best parametric measures, the log area ratio, and the best

	OBJECTIVE QUALITY MEASURES	MAXIMUM SINGLE RESULT		MAXIMUM COMPOSITE RESULT		DISTORTION SUBSET
		$\hat{\rho}$	$\hat{\sigma}_e$	$\hat{\rho}$	$\hat{\sigma}_e$	
1.	1	.63	7.0	.64	6.9	WBD
2.	1,2,3	.63	7.0	.67	6.7	WBD
3.	1,11	.65	6.8	.69	6.6	WBD
4.	1,2,3,6,7,8,9,10,11,12	.65	6.8	.75	6.0	WBD
5.	11,12	.65	6.8	.67	6.7	WBD
6.	1,2,3,15	.72	6.2	.74	6.0	WBD
7.	4,5	.77	5.7	.78	5.7	WFC
8.	13,14	.93	3.3	.93	3.3	WFC
9.	4,5,13,14	.93	3.3			WBD
10.	10,11	.65	6.8	.69	6.6	WBD
11.	8,9,10,11	.65	6.8	.70	6.5	WBD
12.	6,7,10,11	.65	6.8	.70	6.6	WBD

1. Log Spectral Distance
2. Linear Spectral Distance
3. Spectral Distance
4. SNR
5. Short Time SNR
6. Linear Feedback Distance
7. Log Feedback Distance
8. Linear Parcor Distance
9. Log Parcor Distance
10. Linear Area Ratio
11. Log Area Ratio
12. Energy Ratio
13. Frequency Variant SNR
14. Frequency Variant Short Time SNR
15. Frequency Variant Spectral Distance

Table 6.7.1-1. RESULTS OF THE COMPOSITE DISTANCE MEASURE TESTS TO MEASURE MUTUAL INFORMATION AMONG DIFFERENT DISTANCE MEASURES

spectral distance measures contain some separate information, but are really also quite similar.

In studying the parametric measures, we see that the whole parametric set when combined with the whole spectral distance set (recall 6 systems from this group is still all that is involved) a reasonable improvement is obtained. This illustrates a more or less general phenomenom which was observed. That is that often more improvement was obtained by combining a good measure with a bad measure of a vastly different type than from combining two or more similar good measures. Evidently, the better parametric measures are measuring similar information as the spectral distance measures (line 3), and likewise, the better parametric measures contain similar information (line 5). However, when some of the less good parametric measures are included (lines 4, 10, 11, 12), better overall results are obtained.

In the non-frequency-variant noise measures (line 7), the addition of the SNR to the short time SNR adds little. Similarly, in the frequency variant case (line 8), the addition of the frequency variant SNR adds little to the frequency variant short time SNR. In fact, including all these measures together (line 11) adds little to the frequency variant short time SNR.

Finally, it should be noted that the addition of simple spectral distance measures to frequency variant spectral distance measures (line 6) adds little information not available from the frequency variant case above.

6.7.2 Composite Measures for Maximum Correlation

Because the study of the composite measures was a very time consuming task, it was impossible to study a large number of them in detail Basically, the results from all of the correlation studies plus the results from section 6.7.1 were used to guess at what might be good measures. In all, 12 measures without preclassifications and 8 measures with preclassification were studied. Table 6.7.2-1 describes the best of each of these types of measures and shows their results across ALL and WBD for CA, TSQ, and TBQ.

Several points should be made about these results. First, these are maximum obtainable results, and the robustness of these measures has not been tested. Second, the remarkable gain obtained from the preclassified version was almost solely due to the action of the short time frequency variant signal-to-noise ratio measure. However, with these reservations, these results are clearly quite good.

In a real, fieldable system for objective quality testing, it is not clear how close to the limits observed in this study the results would be. However, this was done across a very large data base with many degrees of freedom, and the results here are the best estimates available at this time.

BEST COMPOSITE MEASURE WITH PRECLASSIFICATION

CLASS: SYSTEMS WHICH ARE SIGNAL + NOISE

MEASURE
#

1. SHORT TIME BANDED SNR [$\delta=1$]
2. LOG AREA RATIO [$\alpha=0$; $p=1$]
3. FREQUENCY VARIANT LOG SPECTRAL [$\alpha=0$; $\gamma=1.0$; $p=4$]
4. PARCOR [$\alpha=0$; $p=1$]
5. LINEAR SPECTRAL DISTANCE [$\delta=1$; $\alpha=2$; $\gamma=0$; $p=2$]
6. ENERGY RATIO [$\alpha=0$; $\delta=.25$]

CLASS: ALL OTHER SYSTEMS

MEASURE
#

1. LOG AREA RATIO [$\alpha=0$; $p=1$]
2. FREQUENCY VARIANT SPECTRAL DISTANCE [$\alpha=0$; $\gamma=1.0$; $p=4$]
3. PARCOR [$\alpha=0$; $p=1$]
4. FEEDBACK [$\alpha=0$; $p=1$]
5. ENERGY RATIO [$\alpha=0$; $\delta=.25$]
6. SPECTRAL DISTANCE [$\delta=1$; $\alpha=2$; $\gamma=0$; $p=2$]

RESULTS

	CA		TSQ		TBQ	
	$\hat{\rho}$	$\hat{\sigma}_e$	$\hat{\rho}$	$\hat{\sigma}_e$	$\hat{\rho}$	$\hat{\sigma}_e$
ALL	.89	3.5	.88	4.0	.41	6.1
WBD	.90	3.5	.90	3.9	.32	3.8

BEST COMPOSITE MEASURE WITHOUT PRECLASSIFICATION

1. LOG AREA RATIO [$\alpha=0$; $p=1$]
2. FREQUENCY VARIANT SPECTRAL DISTANCE [$\alpha=0$; $\gamma=1.0$; $p=4$]
3. PARCOR [$\alpha=0$; $p=1$]
4. SPECTRAL DISTANCE [$\delta=1$; $\alpha=2$; $\gamma=0$; $p=2$]
5. ENERGY RATIO [$\alpha=0$; $\delta=.25$]
6. FEEDBACK [$\alpha=0$; $p=1$]

RESULTS

	CA		TSQ		TBQ	
	$\hat{\rho}$	$\hat{\sigma}_e$	$\hat{\rho}$	$\hat{\sigma}_e$	$\hat{\rho}$	$\hat{\sigma}_e$
ALL	.84	4.6	.82	4.9	.38	6.2
WBD	.86	4.2	.86	4.6	.48	6.0

TABLE 6.7.2-1. THE BEST COMPOSITE MEASURES DISCOVERED DURING THIS STUDY

REFERENCES

6.1 P. Mermelstein, "Evaluation of Two ADPCM Codes for Toll Quality Speech Transmission," JASA, to be published, Dec. 1979.